Sample Size Justifications for Pilot Trials of Publicly Funded Randomised Controlled Trials

Amy Lauren Whitehead

Thesis submitted to The University of Sheffield

for the degree of Doctor of Philosophy

School of Health and Related Research

December 2016

The University Of Sheffield.

Access To Thesis.

A fully completed copy of this form must be submitted to Research & Innovation Services prior to the award of your degree. If you are submitting a hard copy of the thesis the form should be bound into the front of the thesis

SECTION 1: STUDENT DETAILS			
Family Name	WHITEHEAD	First Name	AMY
Registration Number	110259785	Department	SCHARR
Thesis Title	SAMPLE SIZE JUSTIFICATIONS FOR PILOT TRIALS OF PUBLICLY FUNDED RANDOMISED CONTROLLED TRIALS		

SECTION 2: THESIS SUBMISSION DETAILS – PLEASE SELECT ONE OF THE FOLLOWING OPTIONS	
☐	I am submitting in print format only for deposit in the University Library **(Note: this option only applies to students who initially registered prior to 2008)**
☑	I am submitting an eThesis only to the White Rose eTheses Online server. I confirm that the eThesis is a complete version of my thesis and no content has been removed
☐	I am submitting an eThesis to the White Rose eTheses Online server and also submitting in print format because I have removed some content from my eThesis

SECTION 3: EMBARGO DETAILS – PLEASE SELECT FROM THE FOLLOWING OPTIONS

Each Faculty has agreed a pre-approved embargo threshold (Arts & Humanities – 1 yr; Engineering – 1 yr; Medicine, Dentistry & Health – 2 yrs; Science – 5 yrs; Social Sciences – 3 yrs. Requests for embargoes that exceed the Faculty threshold will require Faculty approval. If you wish to request a longer embargo, please complete and submit the form available at: www.shef.ac.uk/ris/pgr/code/embargoes

Please note that if no boxes are ticked, you will have consented to your thesis being made available without any restrictions.

Should the thesis be embargoed? If 'Yes', please specify the length of embargo requested (in years)				
	Print Thesis	☐ No	☑ Yes	2 Years
	eThesis	☐ No	☑ Yes	2 Years

Reason for the embargo (please select from the following options):

☐	Third party copyright	☐	Commercial confidentiality
☐	Contains personal data	☐	Could prejudice national security
☐	Could endanger health and safety	☐	Exempt under another category listed in the FOI Act 2000
☑	Planned publication	☐	Other

SECTION 4: COPYRIGHT LICENCE OPTIONS – PLEASE SELECT ONE OF THE FOLLOWING

This thesis is protected by the Copyright Design and Patents Act 1988. No reproduction is permitted without consent of the author. It is recommended that you make your thesis available using a Creative Commons Licence http://creativecommons.org/about/license/. This Licence protects you as the author of the work and also clarifies the uses that others may make of your work.

☑	Creative Commons Attribution-Non-Commercial-No-derivatives **(recommended)**	☐	Creative Commons Attribution-Non-Commercial
☐	Creative Commons Attribution	☐	Creative Commons Attribution-No-derivative-Works
☐	Creative Commons Attribution-Non-Commercial-Share Alike	☐	Other/Do not apply a Licence

SECTION 5: THESIS DEPOSIT AGREEMENT - STUDENT		
1. I, the author, confirm that the Thesis is my own work, and that where materials owned by a third party have been used, copyright clearance has been obtained. I am aware of the University's *Guidance on the Use of Unfair Means* (www.sheffield.ac.uk/ssid/exams/plagiarism).		
2. I confirm that all copies of the Thesis submitted to the University, whether in print or electronic format, are identical in content and correspond with the version of the Thesis upon which the examiners based their recommendation for the award of the degree (unless edited as indicated above).		
3. I agree to the named Thesis being made available in accordance with the conditions specified above.		
4. I give permission to the University of Sheffield to reproduce the print Thesis (where applicable) in digital format, in whole or part, in order to supply single copies for the purpose of research or private study for a non-commercial purpose. I agree that a copy of the eThesis may be supplied to the British Library for inclusion on EThOS and WREO, if the thesis is not subject to an embargo, or if the embargo has been lifted or expired.		
5. I agree that the University of Sheffield's eThesis repository (currently WREO) will make my eThesis (where applicable) available over the internet via an entirely non-exclusive agreement and that, without changing content, WREO and/or the British Library may convert my eThesis to any medium or format for the purpose of future preservation and accessibility.		
6. I agree that the metadata relating to the eThesis (where applicable) will normally appear on both the University's eThesis server (WREO) and the British Library's EThOS service, even if the eThesis is subject to an embargo.		
Student's name (PLEASE PRINT): AMY WHITEHEAD	Signature: Amy Whitehead	Date 23-2-2016

SECTION 6: THESIS DEPOSIT AGREEMENT - SUPERVISOR		
I, the supervisor, agree to the named Thesis being made available in accordance with the conditions specified above.		
Supervisor's name (PLEASE PRINT): Steven A. Julious	Signature: Steven A Julious	Date: 22 February 2016

SECTION 6: TO BE COMPLETED BY RESEARCH & INNOVATION SERVICES		
Does the embargo exceed the agreed Faculty length? FCA - 1YR; FCE - 1YR; FCM - 2YRS; FCP - 5 YRS; FCS - 3 YRS	☐ Yes* if 'yes' please attach embargo extension request form	☐ No
University stamp		

Achievements During PhD

Papers

WHITEHEAD, A. L., JULIOUS, S. A., COOPER, C. & CAMPBELL, M. J. 2015. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research.*

TEARE, M. D., DIMAIRO, M., SHEPHARD, N., HAYMAN, A., WHITEHEAD, A. & WALTERS, S. J. 2014. Sample Size Requirements to Estimate Key Design Parameters from External Pilot Randomised Controlled Trials: A Simulation Study. *Trials,* 15.

LEE, E. C., WHITEHEAD, A. L., JACQUES, R. M. & JULIOUS, S. A. 2014. The Statistical Interpretation of Pilot Trials: Should significance thresholds be reconsidered? *BMC Medical Research Methodology,* 14**,** 41-41.

WHITEHEAD, A. L., SULLY, B. G. O. & CAMPBELL, M. J. 2014. Pilot and Feasibility Studies: Is there a difference from each other and from a randomised controlled trial? *Contemporary Clinical Trials,* 38**,** 130-133.

BILLINGHAM, S. A. M., WHITEHEAD, A. L. & JULIOUS, S. A. 2013. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Med Res Methodol,* 13**,** 104.

Conferences and Presentations

September 2015: Oral Presentation - 'Designing Randomised Controlled Trials Based on Internal Pilot Trials', Royal Statistical Society Conference, The University of Exeter.

August 2015: Oral Presentation and Chair of Sponsors Session - 'Designing Randomised Controlled Trials Based on Internal Pilot Trials', Statistics Research Students Conference, Leeds University.

May 2015: Oral Presentation – 'Designing Randomised Controlled Trials Based on Pilot Trials', Society for Clinical Trials Conference, Washington DC, USA.

September 2014: Poster Presentation – 'The Statistical Interpretation of Pilot Trials: Should Significance Thresholds be Reconsidered?', Royal Statistical Society Conference, Sheffield.

May 2014: Attended – Society for Clinical Trials Conference, Philadelphia, USA.

May 2014: Poster Presentation – 'Setting Pilot Trial Sample Sizes to Minimise Overall Study Costs', Women in Statistics Conference, Research Triangle Park, Raleigh, USA.

April 2014: Oral Presentation and Chair of Session – 'Estimating Sample Sizes for External Pilot Trials Accounting for the Size of the Definitive Study', Statistics Research Students Conference, Nottingham University.

September 2013: Attended – Burwalls Conference for Teachers of Medical Statistics, Oxford University

March 2013: Oral Presentation – 'Statistical Issues in the Design of Pilot Studies for Publicly Funded Randomised Controlled Trials', Statistics Research Students Conference, Lancaster University

Prizes and Nominations

Nominated by students in the School of Health and Related Research for a Teaching and Learning Excellence Award 2015.

Nominated by students for the 'Best Postgraduate Teacher' Award 2015.

Nominated for Student Employee of the Year 2015 for work in the School of Health and Related Research and the Mathematics and Statistics Help Centre.

3rd Place Prize for Oral Presentation at Statistics Research Students Conference, Leeds University, 'Designing Randomised Controlled Trials Based on Internal Pilot Trials'.

Recipient of the Thomas Chalmers Scholarship from the Society for Clinical Trials 2015, Washington DC, USA, 'Designing Randomised Controlled Trials Based on Pilot Trials'.

3rd Place Prize for Poster Presentation at the Royal Statistical Society Conference 2014, Sheffield, 'The Statistical Interpretation of Pilot Trials: Should Significance Thresholds be Reconsidered?'.

Other Activities

2012 – 2016. Graduate Teaching Assistant in the School of Health and Related Research, Sheffield University.

2012 – 2013. Co-supervised a Medical Student Research Project. An Audit of Sample Sizes for Pilot and Feasibility Trials being undertaken in the United Kingdom Registered on the United Kingdom Clinical Research Network Database. Sophie Billingham, Amy Whitehead and Steven Julious.

2013 – 2015. Statistics Tutor, Mathematics and Statistics Help Centre, Sheffield University.

2013 – 2015. Department Postgraduate Representative for the School of Health and Related Research.

2013 – 2015. Member of the Postgraduate Research Student Forum for the Faculty of Medicine, Dentistry and Health.

2013 – 2014. Chair of the Postgraduate Research Student Forum for the Faculty of Medicine, Dentistry and Health.

2013 – 2015. Statistics Tutor for the online MSc Clinical Trials for Edinburgh University.

Summer 2014. Co-supervised a Wellcome Trust Vacation Internship. An Investigation into the Number of Dropouts in Pilot and Definitive Clinical Trials. Edward Pottrill, Amy Whitehead and Cindy Cooper, £2,000.

2015 – 2016. Research Assistant (Maternity Cover) for consultative research with Rotherham General Hospital.

Abstract

A sample size estimate for a clinical trial is an important issue as incorrectly estimating it could have both ethical and financial implications for the trial. Calculating the required sample size for a trial with a continuous outcome requires an estimate of the population variance. A pilot trial can be used to get an estimate of the population variance. However, pilot trials are often small and may give imprecise estimates; adjustment methods are discussed which allow for this imprecision.

Theoretical minimum values for the overall trial sample size when using an adjustment method to design the main trial after an external pilot trial are provided. Using the results recommendations for external pilot trial sample size are presented which aim to minimise the overall trial sample size. It was found that the optimal pilot trial sample size increases with the size of the main trial, therefore stepped rules of thumb are proposed. For a 90% powered main trial this method indicates that the sample size for a two-armed pilot trial to minimise the overall sample size should be 150, 50, 30 and 20 for standardised effect sizes (δ) of $\delta<0.1$, $0.1\leq \delta<0.3$, $0.3\leq \delta<0.7$ and $\delta\geq 0.7$ respectively.

The work is extended to allow for unequal cost per patient between the two trials. The results show that when the pilot trial is less expensive per patient than the main trial the optimal pilot trial sample size increases, giving more precision for the variance estimate and a relatively small main trial. The opposite is true when the main trial is less expensive than the pilot trial. For a 90% powered main trial this method indicates that the sample size for a two-armed pilot trial to minimise the overall sample size should be between 40-260, 20-80, 20-40 and 20-30 dependent on the relative cost of the pilot and main trial per participant for standardised effect sizes (δ) of $\delta<0.1$, $0.1\leq \delta<0.3$, $0.3\leq \delta<0.7$ and $\delta\geq 0.7$ respectively.

For internal pilot trials it is shown that the restricted sample size recalculation procedure raises the average sample size and power of the main trial. Aiming to minimise the overall trial sample size, it was found that the optimal pilot trial sample size rises as the main trial size increases.

The work presented aims to help researchers choose sample sizes for pilot trials and to assess the impact selected methods have on the power and required sample size of the subsequent main trial.

For my Grandma Whitehead who couldn't understand why my 'book' was taking so long.

Acknowledgements

I would like to thank my supervisors Professor Steven Julious, Professor Cindy Cooper and Professor Michael Campbell for the many opportunities they have given me, for their help and support throughout my time in Sheffield, not only with my PhD but also outside of my studies.

I would also like to thank my partner Matthew for his patience, support and for doing all the washing up.

Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	Analysis of Variance
AP	Average Power
ASS	Average Sample Size
BMJ	British Medical Journal
cdf	Cumulative distribution function
CHMP	Committee for Medicinal Products for Human Use
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
df	Degrees of freedom
EME	Efficacy and Mechanism Agency
EMEA	European Medicine Agency
HS&DR	Health Services and Delivery Research
i4i	Invention for Innovation
ICC	Intraclass Correlation
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
IF	Inflation Factor
IQR	Interquartile Range
MCID	Minimum Clinically Important Difference
MRC	Medical Research Council
NCT	Non-Central T-distribution
NETSCC	NHR Evaluation, Trials and Studies Coordinating Centre
NHS	National Health Service
NICE	National Institute for Care and Excellence
NIGMS	National Institute of General Medical Sciences
NIH	National Institutes of Health
NIHR	National Institute for Health Research
NRES	National Research Ethics Service
pdf	Probability distribution function
PDG	Programme Development Grants
PGfAR	Programme Grants for Applied Research
PHR	Public Health Research
RCT	Randomised Controlled Trial
RfPB	Research for Patients Benefit
SD	Standard Deviation
SR	Systematic Review
UCL	Upper Confidence Limit
UK	United Kingdom
UKCRN	United Kingdom Clinical Research Network

Chapter 1

Introduction

Health service research to assess a new health technology can involve a number of investigations including: observational studies, case studies, focus groups, interviews, expert opinions, systematic reviews and clinical trials (Evans, 2003). A clinical trial is defined by Altman (1990) p. 440 as '… a planned experiment on human beings, which is designed to evaluate the effectiveness of one or more forms of treatment.' A series of clinical trials from pilot investigations through to large Randomised Controlled Trials (RCTs) may be conducted in the evaluation of a health technology.

The aims of this thesis are to:

- Provide background information on the area of pilot trials, including definitions, current sample sizes and analysis methods (Chapter 1, 3 and 6)
- Investigate how using an estimate of the variance from a pilot trial to plan a main trial affects the power and sample size of the main trial (Chapter 4 and 7)
- Explore methods of setting a sample size for pilot trials (external and internal) which aim to minimise the overall trial sample size (Chapter 4 and 7)
- Examine how the relative cost of the external pilot versus the main trial affects the sample sizes of the two trials to minimise the overall trial cost (Chapter 5)

Although the aims are highlighted here further justifications and the development of these are presented in Chapters 1, 2, 3 and 6.

This chapter provides an introduction to clinical trials, outlining what they are, how they are funded, the structure of clinical trials in the publicly funded setting and defines associated terminology. Section 1.1 defines the term randomised controlled trial. Section 1.2 describes the funding for clinical trials. Section 1.3 outlines the organisation and structure of publicly funded clinical trials. Section 1.4 compares the terms pilot and feasibility, trial and study in order to outline a definition of the term pilot trial to be used in this thesis. Section 1.5 describes the importance of sample size calculations.

As part of my role as a teaching assistant during the PhD two student research projects were supervised and the results from this work inform the context of the thesis. The first project was carried out by a medical student on a six-week research attachment investigating the current sample size of pilot trials on the United Kingdom Clinical Research Network (UKCRN) database. The results of this work are presented in Section 1.6. The second project was completed by a Wellcome Trust Summer intern, who I supervised on a day-to-day basis. The project looked at how well pilot trials predict parameters for the main trial sample size calculation, specifically predicting the dropout rate and the ratio of randomised to eligible patients. The results of this study are presented in Section 1.7.

There was also additional collaborative complementary research completed during the time of the thesis discussing the available methods for analysing the data from an external pilot trial (Lee et al., 2014) which is presented in Section 1.8. This research adds to the background information of the thesis topic area and is presented here in order to situate the thesis into the wider research area. This chapter is then summarised in Section 1.9 to provide the rationale for the work presented throughout this thesis, the aims of this thesis are also presented in this section. Finally an outline of this thesis is given in Section 1.10.

1.1 Randomised Controlled Trials

In a clinical trial the treatment under investigation, an experimental treatment, is usually compared to a control treatment. The control treatment allows for the fact that patient outcomes can be affected by feelings of enthusiasm, inclusion and expectation. A control treatment should be used to ensure that the effect measured is due to the treatment and not due to the patient's involvement in the clinical trial itself, known as the Hawthorne effect (Parsons, 1974). The control treatment can be an active control, an existing or usual treatment, or a placebo, an inactive treatment made to look exactly the same as the experimental treatment (Pocock, 1983). Using a control treatment is one way to try to reduce the chance of bias in the trial.

Bias is the unconscious distortion in the selection of patients, collection of data, determination of endpoints, and final analyses (Chalmers, 1983). Bias is said to have occurred when there is a systematic difference between the results in the trial and the true value (Petrie and Sabin, 2013). Therefore, bias can give an incorrect estimate of the treatment effect and efforts should be made to lessen its effect (Torgerson and Torgerson, 2008). There are many types of bias, however the main types of bias we are concerned with when designing clinical trials are: the hawthorne effect, allocation bias, assessment bias, confounding, attrition bias and issues with inappropriate analysis methods, for example, multiple testing procedures (Petrie and Sabin, 2013).

Allocation bias occurs when the allocation of patients to treatment is not random and the allocation depends on patient demographics (Pocock, 1983). Randomisation helps to reduce the bias of a trial, there are several types of randomisation including: simple, blocked and stratified (Torgerson and Torgerson, 2008). Randomisation tries to reduce confounding by ensuring that the treatment groups are comparable for any prognostic factors so that any difference between the groups can be attributed to the intervention under investigation (Torgerson and Torgerson, 2008).

Assessment bias occurs when the knowledge of a patients' treatment allocation affects their, the clinicians or other trial staff's behaviour in a way that affects the results of the treatment comparison. Therefore, it is preferable to keep the allocation of treatments to patients hidden to as many trial stakeholders as possible in clinical trials; this is known as blinding (Pocock, 1983).

During a trial participants can withdraw or be lost to follow up, their data can be missing or incomplete. Attrition bias can occur if the pariticpants which dropout of the trial are systematically different from those who remain (Petrie and Sabin, 2013).

A confounder is a variable that is related to both another independent variable and the outcome variable. Ignoring the effect of a confounding variable can distort the association between the independent variable of interest and the outcome variable; perhaps leading to finding a spurious effect of an independent variable on the outcome or missing a true association (Petrie and Sabin, 2013; Daly and Bourke, 2000). Confouding can be controlled by the design or the analysis of the trial (Daly and Bourke, 2000).

Inappropriate analysis methods can lead to bias being introduced into a trial. For example, employing unplanned analyses at the end of a trial. Often subgroup analyses may be conducted at the end of trial without having been previously planned or specified. If sufficient hypothesis tests are carried out eventually a statistically significant difference will be found and the risk of a Type I error increases. If possible all analysis should be pre-specified or multiple-testing procedures should be employed to control the Type I error rate to the appropriate level.

Clinical trials that involve both a control arm and randomisation of patients to treatment are referred to as Randomised Controlled Trials (RCTs). RCTs are seen as the gold standard for testing the effectiveness of an intervention (Torgerson and Torgerson, 2008).

In a clinical trial where the aim is to prove that one treatment is more effective than

another the trial is said to be a superiority trial. Some trials are set up to prove that an intervention is either equivalent or at least not inferior to another intervention. Equivalence trials or non-inferiority trials may be used in situations where the new intervention is say; less toxic, has fewer side effects or maybe costs less than a current treatment (Friedman et al., 2010).

The simplest design for a trial in terms of the comparison of treatment groups is for them to be run in parallel to each other so that if two treatments are under investigation there will be two groups running concurrently each receiving one of the treatments to which patients are randomised. Alternatively another popular design is a crossover trial, where patients receive both treatments consecutively. Here randomisation determines which order each of the trial participants receives the interventions. Other designs include: a paired design, where each participant receives both interventions at the same time and act as their own control; a matched pairs design, where each case is matched to a control across specific confounding variables; a sequential trial, where the results are monitored throughout the trial and the trial is stopped when one treatment is shown to be superior or if it is unlikely a difference will emerge; a factorial design, allows us to investigate the effect of two treatments individually compared to control, compared to each other and the effect of the treatments when used in combination; or an adaptive design, the definition of an adaptive design is discussed in Chapter 6.

1.2 Funding for Clinical Trials

Research funding for clinical trials can broadly be defined as coming from: industry (e.g. Pharmaceutical companies), government (e.g.The Medical Research Council (MRC) or the National Institute for Health Research (NIHR)) or charities. The funding streams may not be mutually exclusive (i.e. some trials are funded by industry and public money).

There are two main public funding bodies for health research in the UK, the MRC and the NIHR. Within the NIHR there are four sections which fund or have funded pilot and feasibility studies:

- The Research for Patient Benefit (RfPB) Programme,
- Programme Grants for Applied Research (PGfAR),
- Programme Development Grants (PDG) and
- The NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC) (NIHR, 2012a).

The RfPB is interested in regionally derived research with clear potential to benefit patients, offer value for money and increase National Health Service (NHS) effectiveness. PGfAR funds priority research for the NHS with potentially a short time until the results will have a practical application. PDG are for preparatory work prior to funding from the PGfAR. Programmes managed by the NETSCC include:

- The Health Technology Assessment (HTA) Programme,
- The Public Health Research (PHR) Programme,
- The Efficacy and Mechanism Evaluation (EME) Programme and
- Health Services and Delivery Research (HS&DR) (NIHR, 2012d).

The HTA programme funds 'health technology' trials defined as 'any method used to promote health, prevent and treat disease and improve rehabilitation and long-term care' (NIHR, 2012a). The PHR funds pilot and feasibility trials evaluating public health interventions outside the NHS. The EME programme funds pilot and feasibility studies to support trials through the early phases. The HS&DR funds a broad range of research assessing quality, and efficiency of the NHS. The Systematic Review (SR) programme and the Invention for Innovation (i4i) Programme are also funded by the NIHR but they do not support pilot trials although i4i may fund a feasibility study as defined by the NIHR (NIHR, 2012a).

A large amount of public money is invested into health research every year. In 2013/14 the UK government invested £1,858M. As previously discussed the investment is made through two main authorities: The MRC (£845.3 million in 2013/14) (MRC, 2015) and the NIHR (£1,013.6 million in 2013/14) (NIHR, 2015). The MRC and NIHR fund research through research facilities, research centres in partnership with universities and by funding researchers (MRC, 2012a). In 2014/15 the MRC employed 2,560 people and provided £62.9 million in scholarships and fellowships funding for 1,440 postgraduate students and 390 fellows (MRC, 2015). Due to the large amounts of public money involved, careful stewardship is required by these authorities to ensure that the money is invested appropriately and will be used in the most effective manner in order to improve clinical practice and bring health benefits to the population (MRC, 2012b).

1.3 Structure of Publicly Funded Trials

In the private sector a portfolio of evidence for an intervention is built up over phases of development (Pocock, 1983):

- Phase I – performed in healthy volunteers, assessing safety and tolerability
- Phase II – in patient, first assessment of efficacy and dose finding trials
- Phase III – in patient, pivotal RCT programme
- Phase IV – post-marketing surveillance (Julious et al., 2010).

In the public sector people tend to either launch straight into the main definitive RCT or conduct a pilot trial beforehand (McDonald et al., 2006). A reason for this difference could be that usually interventions tested in publicly funded research are, drugs which are already licensed, a health technology or complex interventions; whereas in the privately funded setting (industry) the focus is likely to be on unlicensed drug trials of new chemical compounds untested in a clinical setting for safety and efficacy. A health technology is 'the application of organised knowledge and skills in the form of devices, medicines, vaccines, procedures and systems developed to solve a health problem and improve quality of lives' (WHO, 2015).

Some examples of health technology trials that have taken place at the University of Sheffield are:

- CoSMoS: a trial looking at computerised cognitive behavioural therapy for the treatment of depression in patients with multiple sclerosis (Cooper et al., 2011).
- RATPAC: a trial comparing a new diagnostic strategy for suspected myocardial infarction patients against current practice (Goodacre et al., 2010).
- CACTUS: a trial looking at computer based self managed word finding therapy for people who experience aphasia (a communication disorder) after stroke (Palmer et al., 2011).

Complex interventions are 'made up of various interconnecting parts' (Campbell et al., 2000) which may act independently or inter-dependently; it is therefore difficult to distinguish any specific active ingredient (MRC, 2000). Examples of complex interventions that have taken place at the University of Sheffield include:

- Booster: a trial looking at the effectiveness of motivational interviewing techniques to promote and sustain change in physical activity in middle-age adults in deprived urban areas. Participants were given a DVD which used motivational interviewing to promote increased physical activity, after this they were randomised to either control or one of two interventions: motivational interviewing techniques via telephone consultations or face-to-face meetings (Hind et al., 2010).
- Lifestyle Matters: a trial determining the benefit of an occupational therapy programme for people aged 65 and over. The programme enables the participants to become more active, engage in the community and make independent choices, thus increasing their health and wellbeing; which is thought to be strongly associated with good mental wellbeing (Sprange et al., 2013).

Pilot trials are particularly important for trials of complex interventions compared to drug trials. Trials of complex interventions may require long-term commitment from the

participant or a major change in lifestyle; thus a good prediction of adherence and dropout rates is needed to test if the trial is feasible (MRC, 2000). In addition, due to the complicated nature and number of trial processes, it is especially important to test the methods that will be used in the main RCT e.g. methods of recruitment, randomisation, follow-up etc. as well as obtain estimates in order to calculate the sample size required for the main trial (MRC, 2000).

It could be argued that thorough pilot testing is particularly important in publicly funded trials as it is important to reduce the amount of public money wasted due to bad planning or avoidable mistakes in the main trial. For these reasons this thesis focuses on clinical trials in the publicly funded setting.

1.4 Pilot and Feasibility Trials

There is much discussion in the literature surrounding the definition of the term pilot as well as another term also used to describe preliminary work, feasibility. This section compares the two terms, their similarities and differences, as well as other terms used to describe preliminary studies, before concluding with a definition to be used in this thesis.

1.4.1 Definitions from the Literature

A pilot study is defined by Thabane et al. (2010) as a piece of work conducted before a more substantive trial to aid in the development of the design for the future trial. Pilot studies should assess methodological issues with the trial design rather than concentrate on the efficacy or effectiveness of the treatment.

The NETSCC (NIHR Evaluation, Trials and Studies Coordinating Centre), which is responsible for managing evaluation research for the NIHR, define a pilot study as 'a version of the main study run in miniature to tell whether the components of the main

study can all work together'. They suggest that a pilot should focus on how the trial runs (i.e. the recruitment process, the randomisation procedures, and the treatment and follow up assessments) (NETSCC, 2012). There must also be a plan for further work. This definition is comparable to the UK NICE (National Institute for Care and Excellence) definition of a pilot study as 'a small scale "test" of a particular approach ... the aim would be to highlight any problems or areas of concern and amend it before the full scale study begins' (NICE, 2013). The plan for future work is crucial for pilot studies otherwise they could be seen as underpowered trials, which are unethical and have limited scientific use.

The National Institutes of Health (NIH) is the main government organisation in the United States that is responsible for medical research. Within the NIH sits the National Institute of General Medical Sciences (NIGMS) (NIGMS, 2012). NIGMS defines pilot studies as 'trial runs designed to improve data collection instruments and procedures before the data collection effort is begun' (NIGMS, 2011).

Lancaster et al. (2004) highlight seven examples of reasons for conducting a pilot study:

1. To guide a later sample size calculation
2. To test the integrity of study protocol
3. To test data collection forms or questionnaires
4. To test the randomisation procedure
5. To investigate the recruitment and consent rates
6. To investigate the acceptability of the intervention
7. To select the most appropriate outcome measure

Thabane et al. (2010) add three more objectives to this list:

8. To assess time and budgeting issues
9. To provide data management and staff training
10. To assess the safety and determine dose levels

The aim of the pilot is therefore to provide experience in the running of the trial and to highlight any problems, so they may be corrected before the main trial begins (Teijlingen and Hundley, 2001).

The NETSCC also define a type of study called a feasibility study which they define as 'studies used to estimate important parameters that are needed to design the main study'. For example this could be, the standard deviation of the outcome measure needed to complete the sample size calculation for the main trial, the number of people eligible or response rates in order to assess whether it will be possible to meet the sample size requirement (NETSCC, 2012). NIGMS offer a slightly different definition of a feasibility study, 'a preliminary evaluation aimed at determining the optimal approach for a full scale outcome evaluation' (NIGMS, 2011). Therefore, the aim of a feasibility study is to assess whether it is possible to perform a full main trial.

A review of papers published in 2004 of seven major journals looked at the objectives of pilot studies in the literature to clarify the definition of a pilot study (Lancaster et al., 2004). This work was extended in 2010 to distinguish between pilot and feasibility studies (Arain et al., 2010). Arain et. al (2010) report that studies labelled pilot tend to use stricter methodology (i.e. a sample size calculation, randomisation and blinding) and are more likely to conclude the need for further work than studies labelled feasibility. This seems to suggest that there is some distinction between the usage of the two terms in practice. These authors recommend the use of the NETSCC definitions.

Although some authors agree that there is a distinction between a pilot and a feasibility study (Arain et al., 2010, Lancaster et al., 2004) several do not distinguish between these two terms specifically (Arnold et al., 2009, Thabane et al., 2010). Thabane et al. (2010) state that the two terms are equivalent, while Arnold et al. (2009) argue that the term feasibility does not reflect the scope of many pilot studies. Arnold et al. (2009) instead use three separate terms to describe the development stages of a trial. They define pilot work as 'any background research that informs a future study'; a pilot study as ‘a study with a

specific hypothesis, objective and methodology' and a pilot trial as 'a stand-alone pilot study with a randomisation procedure'.

During the PhD I embarked on a project with colleagues to look at the definitions of pilot and feasibility studies and consider whether there is a difference from each other and from a randomised controlled trial. Following the review of the literature presented above and studying example trials, we formed an idea that the term feasibility should be used as an overarching term for preliminary studies and the term pilot refers to a specific type of study that mimics the definitive trial design (Whitehead, Sully and Campbell, 2014). However, a pilot trial although mimicking the definitive trial design should have feasibility aims if not, it could be perceived as a small underpowered study with limited scientific use and may be deemed unethical.

This was later echoed by the work of Eldridge et al. (2016) who worked to clarify the definition of pilot and feasibility while working on an extension of the CONSORT guidelines for pilot and feasibility studies. They conducted a large Delphi study (N=93 in round 1 and N=79 in round 2) and an international expert consensus meeting as well as a systematic review. During the review they found that it was not possible to apply mutually exclusive definitions of pilot and feasibility studies that are consistent with the way authors describe their studies. This supports the idea that the terms feasibility and pilot are not necessarily separate concepts.

They conclude that feasibility is thus an overarching concept within which they describe three distinct types of study: randomised pilot studies, non-randomised pilot studies and feasibility studies (Eldridge et al., 2016).

1.4.2 Definition for Thesis

From the review of the literature several conclusions can be drawn: a pilot study investigates not only the trial processes and procedures to avoid problems in the later

trial but also the feasibility of that later main trial. The aim of a pilot study is not to assess the superiority of one treatment over the other but can be any of the objectives listed by Lancaster et al. (2004) or Thabane et al. (2010) amongst others. By comparison feasibility is an overarching term describing many types of preliminary work and pilot studies are a special type of feasibility study that mimic the design of the main trial.

The definitions put forward by Arnold et al. (2009) describe the flow of work through the whole clinical trial process well, therefore to incorporate these into the definition to be used in this thesis, this thesis defines a pilot trial as a pilot study (as described above), which also involves randomisation between a control and experimental treatment group.

1.5 The Importance of Sample Size

Despite the thorough and meticulous application procedure for research funding, most trials require extensions which cost more money and delay the use of findings in clinical practice (McDonald et al., 2006). In 2006, McDonald et al. carried out a review of 114 trials from the UK MRC and NIHR HTA programme between 1994 and 2002. They found that only 55% of reviewed trials recruited to within 80% of the original target. Additionally, 54% of trials requested an extension to the trial grant in order to complete the original trial. In 14 (11%) of the trials enrolment was stopped before the planned end of the recruitment period. Of these trials the decision to stop early was due to poor recruitment in 11 of them.

This review was updated in 2013 investigating trials which recruited participants between 2002 and 2008, 73 trials were included in the analysis (Sully et al., 2013). Of these trials 55% recruited to their original sample size target and 78% recruited 80% of their target sample size. This is an improvement over the figures seen in the previous review. However, extensions were still common with 45% of trials receiving an extension (Sully et al., 2013). These findings are reflected throughout the literature (Relton et al., 2010, Watson and Torgerson, 2006). Of multicentre trials published in the BMJ and the Lancet

between 2000 and 2001, 51% reported problems with recruitment these included; competing trials, problems with ethics, inaccurate incidence information, clinician resistance and a narrowly defined trial population (Puffer and Torgerson, 2003).

If a trial sample size is not large enough, results may be less accurate, the power to find a difference will be reduced, and the trial results may be inconclusive (Altman, 1980). It is clear that a large amount of trials fail not due to lack of treatment efficacy but because of poor trial design and unexpected low recruitment rates. The Cooksey Review (2006) commissioned by Gordon Brown (the Chancellor of the Exchequer at the time) aimed to look at the best institutional arrangement for the new single fund for health research, the findings clearly reflect the conclusions of McDonald et al. (2006) and Sully et al. (2013) stating that; more work was needed to ensure that publicly funded health research is carried out in the most effective and efficient way (Cooksey, 2006).

With large amounts of public money being invested every year, it is important for investigators who have been allocated funding to ensure the best use of the money is being made. The McDonald review showed over half of the trials which conducted a pilot study made changes to their recruitment strategy, demonstrating how pilot trials may assist in the design of the main trial.

Having an accurate sample size calculation is important for effective allocation of public resources as well as ethical conduct. The results of any trial are subject to the possibility of error. At the end of a trial we could conclude that there is a difference between the two groups under investigation when in fact there is not; or we could conclude that there is no difference when in fact there is. A sample size calculation helps us to compute the minimum number of people we need to reduce the risk of these errors to a level we are willing to accept (Kirkwood and Sterne, 2003).

In order to perform a sample size calculation for a trial with a continuous outcome variable the investigator needs to have estimates of:

- the acceptable level of Type I error,
- the acceptable level of Type II error,
- the allocation ratio of patients to the experimental and control group,
- the variance of the primary outcome measure and,
- the Minimum Clinically Important Difference (MCID).

A Type I error is the probability of rejecting the null hypothesis when it is in fact true. In a superiority trial this would lead to the conclusion that there was a difference between the two treatment groups when in fact no difference exists in the true underlying population. A Type II error is the probability of not rejecting the null hypothesis when it is in fact false. In a superiority trial this would lead to a conclusion that there is no difference between the two treatment groups when in fact there is a difference in the overall population (Petrie and Sabin, 2013). These errors will be discussed further in Chapter 2. The variance is a measure of how spread out the data are from the mean value. The larger the variance the more spread out the data are (Petrie and Sabin, 2013). The MCID is:

> The smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management (Jaeschke et al., 1989, p.408).

More simply, the minimum clinically important difference is the difference in the treatment response between the groups, which would cause a change in clinical practice. Although there are many methods available to formulate an estimate of the MCID (Rai et al., 2015) this value can be based on expert opinion. There are traditionally accepted levels of the Type I and II errors, 5% for Type I error and either 10 or 20% for the Type II error (Julious, 2009). However, when it comes to estimating the variance data from past trials or a pilot trial must be used.

A sample size justification is an important consideration in the planning of a clinical trial (Julious, 2009, Machin et al., 2008). Journals for example, the British Medical Journal (BMJ) (BMJ, 2012) and the Lancet (Lancet, 2012), require that any submission should follow the Consolidated Standards of Reporting Trials (CONSORT) guidelines (CONSORT, 2012). The CONSORT guideline, point 7a, states that authors should provide details of how the sample size was determined (CONSORT, 2012). In addition ethical committees, for example the National Research Ethics Service (NRES), require a sample size justification (NRES, 2012) as well as funding bodies such as the NIHR RfPB (RfPB, 2012) and the HTA programme (HTA, 2012). The MRC (MRC, 2012c) in fact states that the requested sample structure and size should be sufficient to generate meaningful results, suggesting a power calculation could be required. All trials require a sample size justification but it may not always be practical to do a sample size calculation, this will be further discussed in Chapter 3.

It is important to consider why funders, journals and ethical committees require a sample size justification. Trials that are too small or too large can cause ethical issues (Altman, 1980). For example, if there are too few participants you may not be able to answer the research question being investigated (Campbell et al., 2010). If the sample size is too small the probability that the trial will find a statistically significant result even if one exists (the power of the trial) will be low, therefore, patients are involved in a trial, which has little chance of having scientific relevance and validity. Secondly, due to the small sample size the confidence in the estimate of the interventions effect may be so wide that a difference of neither zero nor the MCID can be ruled out hence the results would be inconclusive (Halpern et al., 2002). Conversely, if the sample size is too large resources could be being wasted, such that more patients than necessary may be given a treatment, which will later be shown to be inferior or result in a delay in effective treatment being released on to the market (Altman, 1980).

Routinely conducting pilot trials could reduce the number of definitive trial failures and extensions by helping to identify problems early and therefore save money in the long run. More accurate calculation of the required sample size of a trial can help to stop the waste of resources that is caused by either inadequate sample size to fulfil the trial aims or over recruitment of patients to trials. The focus of this thesis will be on using pilot trials to estimate the required sample size of the main trial.

1.6 Current Sample Sizes of Pilot Trials

Section 1.4 looked at the definition of a pilot trial and the reasons why one might be conducted. This section will focus on how many people are typically recruited to a pilot trial (this topic will also be discussed in more detail in Chapters 2, 3 and 4). To address this question this section presents an audit, which investigated the current sizes of pilot and feasibility trials registered on the United Kingdom Clinical Research Network (UKCRN) database (Billingham et al., 2013). This project was carried out by a medical student whom I co-supervised during my PhD, the paper written from the project is presented in Appendix E.

The UKCRN database was chosen for this audit as it contains the NIHR portfolio of England and the corresponding portfolios for Northern Ireland, Scotland and Wales. It is not compulsory for a trial to register on this database, however in doing so a trial receives the support of the clinical research network.

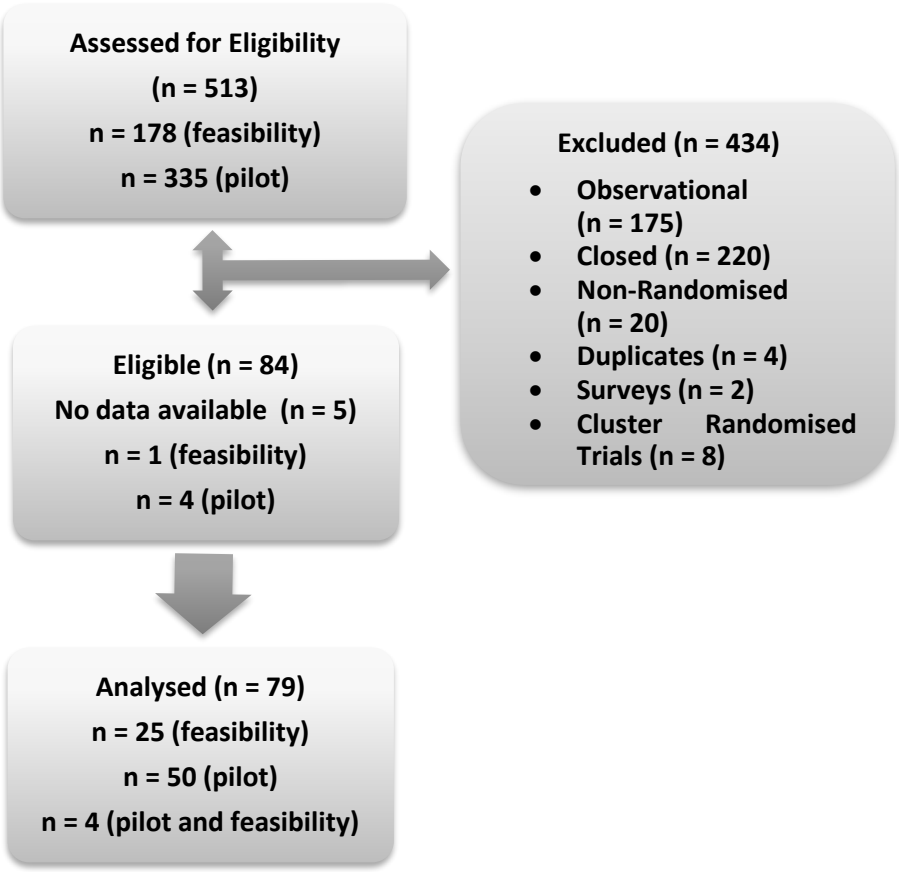
The aim of the study was to investigate the sample size of ongoing pilot/ feasibility trials in the UK. The decision was made to search for both pilot and feasibility trial due to the results of the review presented in Section 1.4, when it was found that there are some differences between the two types of investigation.

The UKCRN database was searched using the keywords 'pilot' or 'feasibility'. The inclusion criteria for this study were that trials were:

- Randomised controlled trials,
- Currently recruiting participants,
- Interventional,
- Not based on healthy volunteers,
- Not cluster randomised (as these tend to be large compared to individually randomised trials).

The search yielded 513 trials, after the eligibility criteria were applied. Duplicates or trials with no data or contact information were removed, yielding a total of 79 trials left in the study. Of these 79 trials, 25 were labelled as feasibility, 50 were labelled pilot and 4 were labelled both pilot and feasibility by the investigators. This heterogenity reflects the review in Section 1.4 that while most researchers distinguish between the two terms there are researchers who believe them to be equivalent. The CONSORT diagram for this trial can be seen in Figure 1.1.

Figure 1.1: Flow of Trials Through the Review



Of the trials analysed (n = 79), most of the trials had two treatment groups (86.1%), the majority of trials were of a health technology (75.9%), most of the trials in the review were from the public sector (59.5%) and 57.0% of the trials had a continuous outcome measure as their primary endpoint. These results can be seen in Table 1.1.

Table 1.1: Characteristics of the Trial in the Review

Characteristic		N (%)
Number of Arms	**Two**	68 (86.1%)
	Three	10 (12.7%)
	Four	1 (1.3%)
Type of Trial	**Health Technology**	60 (75.9%)
	Drug	19 (24.1%)
Type of Endpoint	**Dichotomous**	31 (39.2%)
	Continuous	45 (57.0%)
	Time-to-Event	1 (1.3%)
	Other	2 (2.5%)
Funder	**Industry**	13 (16.5%)
	Public	47 (59.5%)
	Charity	19 (24.1%)

Looking at the sample sizes of these studies, the studies labelled pilot had a median sample size of 30 (IQR: 20-45) participants per arm, for feasibility studies the median sample size per arm was 36 (IQR: 25-50). On average the publicly funded trials were larger than industry trials with a median of 36 (IQR: 25-60) compared to 30 (IQR: 16-31) participants per arm respectively. The results also show that trials with a dichotomous outcome were larger on average than the trials with a continuous outcome measure, with 36 (IQR: 25-50) compared to 30 (IQR: 20-50) patients per arm respectively. For the publicly funded pilot trials the median sample size per arm was 36 (IQR: 30-42) and 30 (IQR: 20-60) for dichotomous and continuous outcomes respectively. These results along with the result for feasibility trials are presented in Table 1.2.

Table 1.2: Median Sample Sizes per Arm for Publicly Funded Trials

Trial Type	Type of Endpoint	N	Median (IQR)
Pilot	**Dichotomous**	6	36 (30-42)
	Continuous	21	30 (20-60)
Feasibility	**Dichotomous**	9	50 (30-70)
	Continuous	6	43 (15-60)

The results of this review are relevant to this thesis as they show that most clinical trials have two arms, are investigating health technologies and are based on a continuous outcome measure. For the publicly funded trials it also showed that current pilot trial sample sizes are on average around 30 people per arm for trials where the primary endpoint was a continuous outcome measure, this idea will be revisited in Chapter 4.

1.7 How Predictive of Main Trials are Pilot Trials

In this chapter it has been established that prior to the main trial being conducted a pilot trial may be carried out for multiple reasons, one of which is to help with parameter predictions for a main trial sample size calculation. Once the initial sample size calculation has been completed to give the required number of evaluable patients needed at the end of the trial, several adjustments need to be made to this figure.

When a clinical trial is being set up not everyone who is eligible to be entered into the trial will end up being randomised. In addition, some patients entered into the trial will dropout before the trial completes. In order to know how many people should be approached to be in the trial we also need predictions of the proportion of patients eligible that will be randomised and the proportion that are randomised who will provide evaluable data. This section presents an audit of main trials and the pilot trials which precede them, which investigated how accurately pilot trials predict these quantities for the main trial.

The study aimed to look at the differences between the dropout rates and the ratio of randomised to eligible patients of pilot trials and their respective main trials. An audit of main trials with external pilot trials was carried out in two phases. The first phase was carried out as part of another PhD project (Knox et al., 2014) and the second part was carried out as part of a Wellcome Trust Summer Internship, which I helped to co-supervise during my PhD.

Firstly, publicly funded RCTs published between 2004 and 2013 were collected from HTA monographs (Knox et al., 2014). HTA monographs are the clinical reports of trials that are funded by the HTA. The criteria for inclusion in this part of the study were single or multi-centre randomised controlled trials that were:

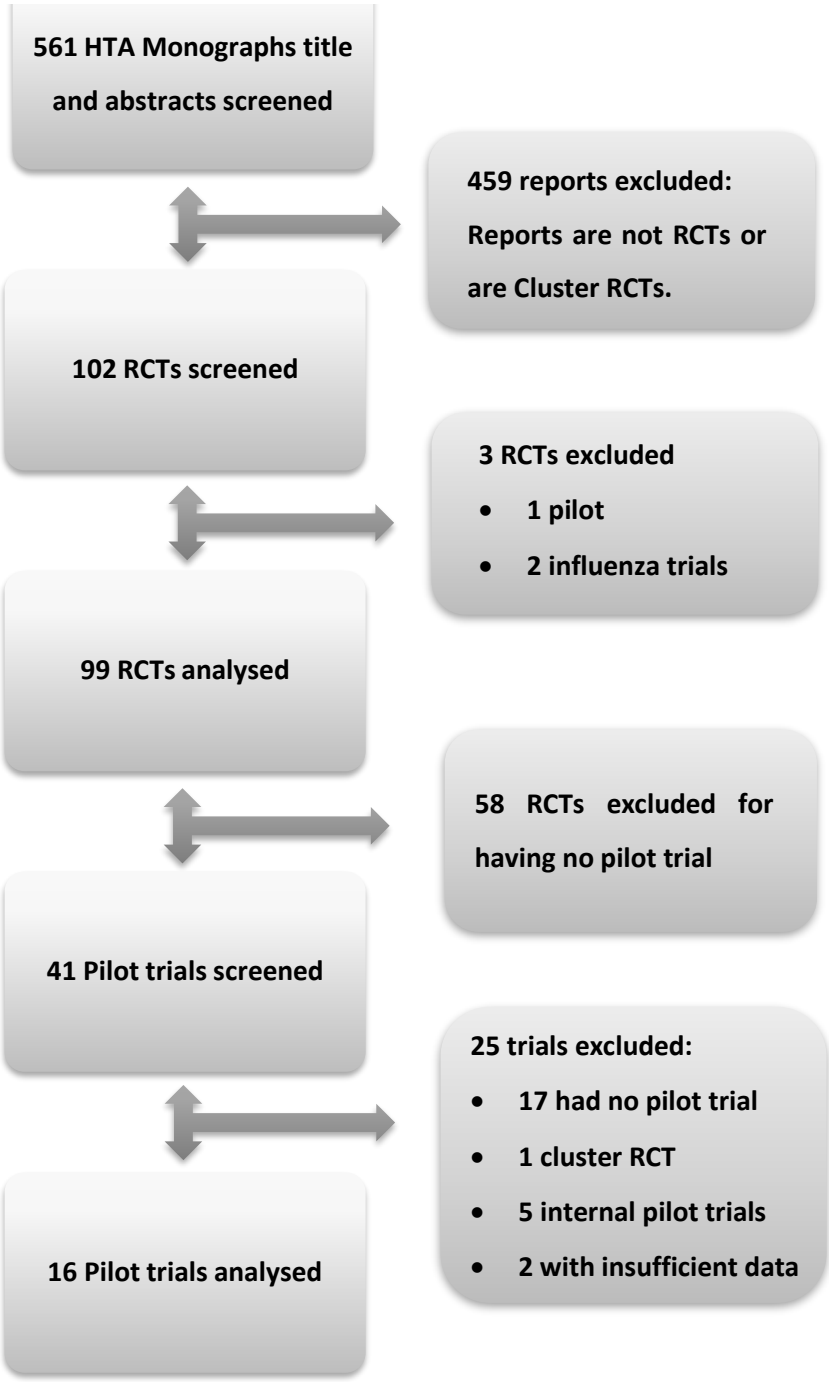
- Not stopped early,
- Not an adaptive design (adaptive designs allow changes to be made to the design or statistical procedures of ongoing clinical trials) (Chow and Chang, 2008),
- Not cluster randomised,
- Not an influenza trial,
- Not a pilot trial.

Trials that had a pilot trial then went forward to the next phase of data collection. From this initial data collection 561 HTA monographs were identified, 99 of which met the inclusion criteria. Of these trials , forty were judged to have a pilot trial at this stage, by reviewing the reports for references to a pilot trial.

During the second phase these 41 RCTs and pilot trials were re-investigated; 17 were reassessed to not having a pilot trial, 5 had an internal pilot but no external pilot and 2 had insufficient data. This left 16 trials where there was information on the main trial and the external pilot trial. Data was extracted on the dropout rate in the main trial and the pilot trial, the number of patients, which were eligible, and the number of patients

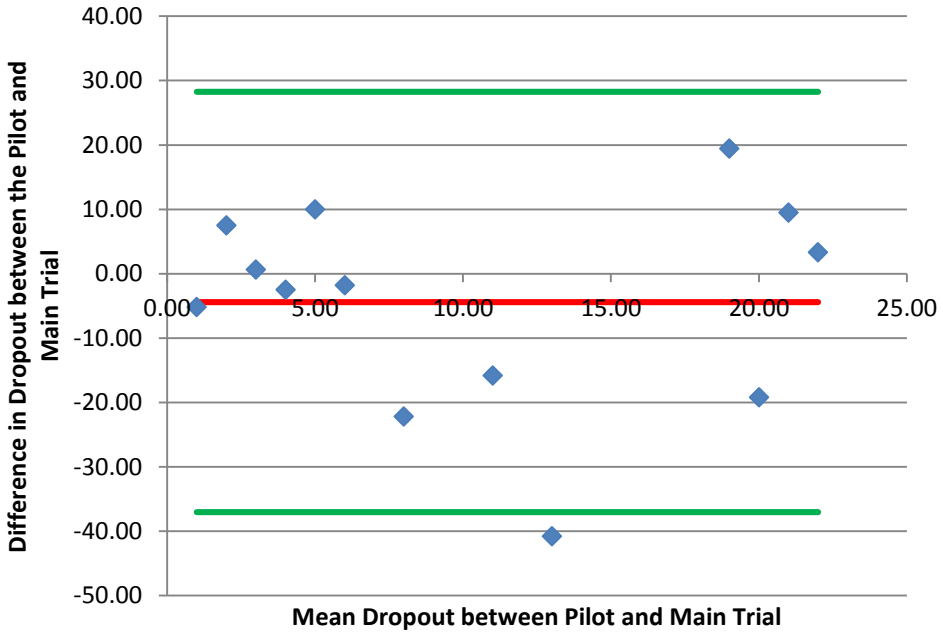
eventually randomised in each trial. The CONSORT diagram showing the flow of trials through this study can be seen in Figure 1.2.

Figure 1.2: Flow of Trials through Phases 1 and 2 of Study



There were 13 trials which had data for both the dropout rate in the pilot and the main trial. The median dropout rate in the pilot trials was 23.48% (IQR = 5.90-31.90). The median dropout rate in the main trials was was 10.79% (IQR = 5.84-28.04). Minimal bias was found as the average difference between the dropout rate in the pilot and the main trial was 4.40%, where the average dropout rate in the main trial was 4.40 percentage points (SD = 16.32) less than in the pilot trial. This can be seen in the Bland Altman plot displayed in Figure 1.3 and in Table 1.3.

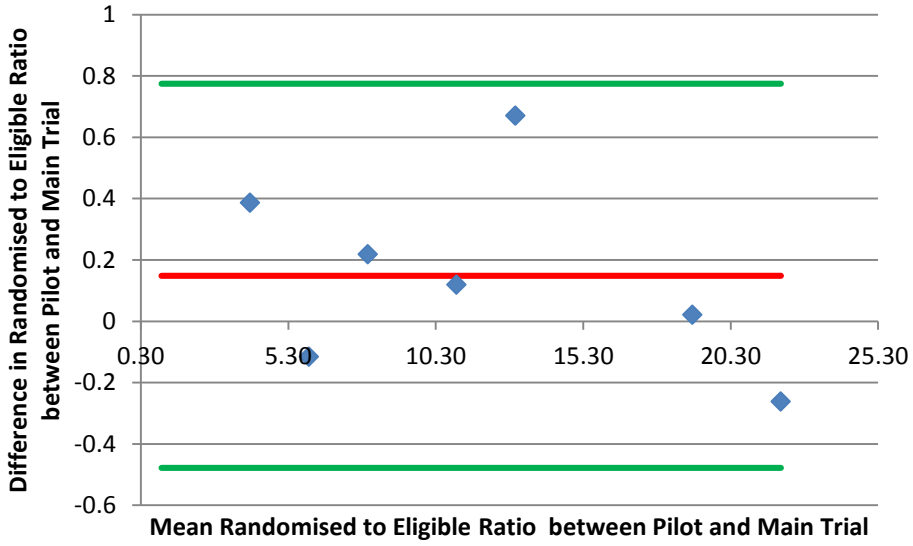
Figure 1.3: Bland-Altman Plot Comparing Percentage Dropout in the Pilot and Main Trial



The median ratio of randomised to eligible patients for the pilot trials was 48.94% (IQR = 30.26 – 61.32%). The median ratio for the main trials was 60.90% (IQR = 45.74 – 92.47%). Seven trials in the study had information on the ratio of randomised to eligible patients for both the pilot and the main trial. The mean of the differences between the two ratios was 14.86% (SD = 31.33) where the ratio of patients randomised to patients eligible is higher in the main trial than the pilot. This shows that on average main trials in the study had a higher rate of converting eligible patients into randomised patients than

the corresponding pilot trial. These results are displayed in the plot shown in Figure 1.4 and Table 1.3.

Figure 1.4: Bland-Altman Plot Comparing Randomised to Eligible Ratios between the Pilot and Main Trial



Also of interest to the research of this thesis is the ability of pilot trials to predict the variance in the main trial. Only three trials had measurements of standard deviation for both the pilot and the main trial. Although the sample size is very small an initial investigation of how similar the standard deviation is in the pilot and the main trial showed that there was very little bias overall. The median in the pilot was 8.74 (IQR = 1.70, 19.30). The median standard deviation in the main trial was 3.64 (IQR = 1.39, 23.80). The mean of the differences between the standard deviations of pilot and main trials was -0.30 (SD = 4.80) where the standard deviations were slightly higher in the pilot trial compared to the main trial. These results can be seen in Figure 1.5 and Table 1.3.

On average there is minimal bias in the prediction of the standard deviation in the main trial from the pilot trial. However the results are highly variable and only have data from

three trials. This issue of the imprecision of the variance estimate from a pilot trial will be further considered and discussed throughout this thesis.

Figure 1.5: Bland-Altman Plot Comparing Standard Deviation between the Pilot and the Main Trial

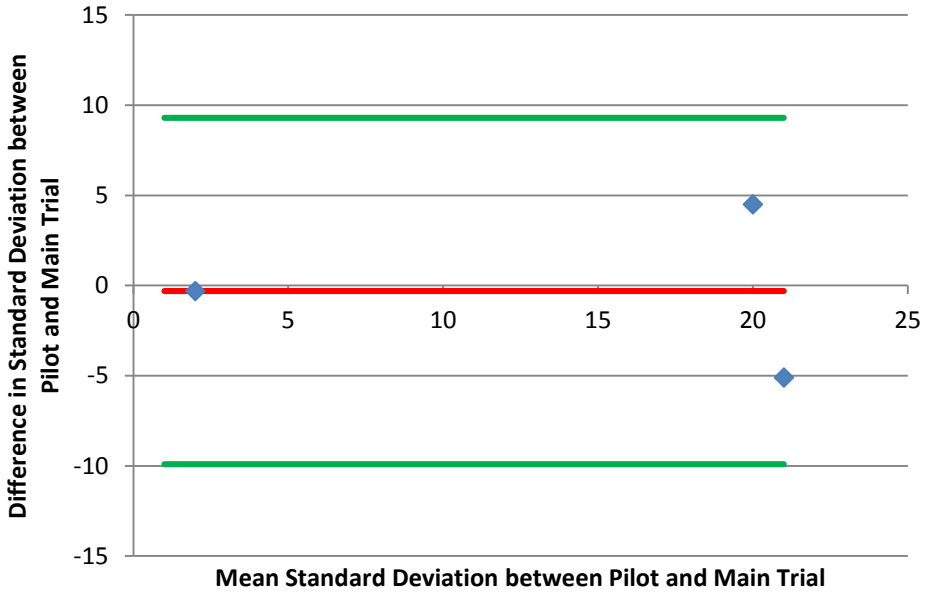


Table 1.3: Results Comparing the Pilot to the Main Trial

Parameter	N	Pilot	Main	Difference
Dropout Rate	13	Mean = 21.12 SD = 15.99 Median = 23.48 IQR = (5.90, 31.90)	Mean = 16.72 SD = 11.66 Median = 10.79 IQR = (5.84, 28.04)	Mean = -4.40 SD = 16.32 Median = -1.80 IQR =(-17.52, 8.49)
Randomised to Eligible Ratio	7	Mean = 50.36 SD = 22.39 Median = 48.94 IQR = (30.26, 61.32)	Mean = 65.22 SD = 23.24 Median = 60.90 IQR = (45.74, 92.47)	Mean = 14.86 SD = 31.33 Median =11.96 IQR = (-11.54, 38.68)
Standard Deviation	3	Mean = 9.91 SD = 8.86 Median = 8.74 IQR = (1.70, 19.30)	Mean = 9.61 SD = 12.34 Median = 3.64 IQR =(1.39, 23.80)	Mean = -0.30 SD = 4.80 Median = -0.31 IQR = (-5.10, 4.50)

1.8 Analysing Pilot Trials

Traditionally during the analysis of a trial a hypothesis test is undertaken such that a P-value is calculated and compared to the 5% significance level to determine whether the difference between the treatments is statistically significant. If the P-value is above 0.05 we do not reject the null hypothesis; if the P-value is less than 0.05 we reject the null hypothesis at the 5% level (hypothesis testing is discussed further in Chapter 2) (Pocock, 1983).

Some authors have suggested using a hypothesis test for the analysis of a pilot trial but with the significance level raised. For example, using a Type I error rate of 0.2 (Stallard et al., 2005) or 0.25 (Schoenfeld, 1980), these will be discussed further in Chapter 3.

Authors seem to be willing to accept a higher probability of Type I errors in pilot trials than main trials. Type I errors have different implications in pilot trials compared to main trials. In the main trial a Type I error would result (for a superiority trial) in an incorrect conclusion of experimental treatment superiority over the control treatment. In a pilot trial a Type I error would lead to a main trial being carried out unnecessarily, however, this mistake can be corrected at the main trial stage (Stallard et al., 2005).

For pilot trials it has been suggested that it may be more informative to present the treatment effect and a range of possible responses (a Confidence Interval (CI)) (Lancaster et al., 2004, Thabane et al., 2010, Julious and Patterson, 2004). This CI need not be a conventional 95% CI and other widths may be considered (Lee et al., 2014). The CI should be interpreted in reference to the MCID and the line of no effect. If the CI contains both the MCID and the line of no effect then either is possible, there could be no difference between the treatments or there could be a difference larger than the MCID (Lee et al., 2014).

This procedure is demonstrated in the work published in Lee et al. (2014), (contibuted to during my PhD, journal paper presented in Appendix E) for a trial comparing the effects of two different methods of treating leg ulcers on the General Health (GH) dimension of the SF-36 questionnaire (The SF-36 is a 36 question tool measuring several quality of life dimensions including: physical functioning, role limitations due to physical health, role limitations due to emotional problems, energy/fatigue, emotional wellbeing, social functioning, pain, general health and perceived change in health) (RAND, 2015). The intervention was clinic based four layer compression bandaging and the control was usual care provided by district nurses at home. The trial recruited 233 patients with venous leg ulcers, 120 to the intervention group and 113 to the control group. The paper used the first 40 patients to mimic a pilot trial, 31 of which had complete data for 3-month SF-36 GH dimension (17 in intervention group and 14 in control group).

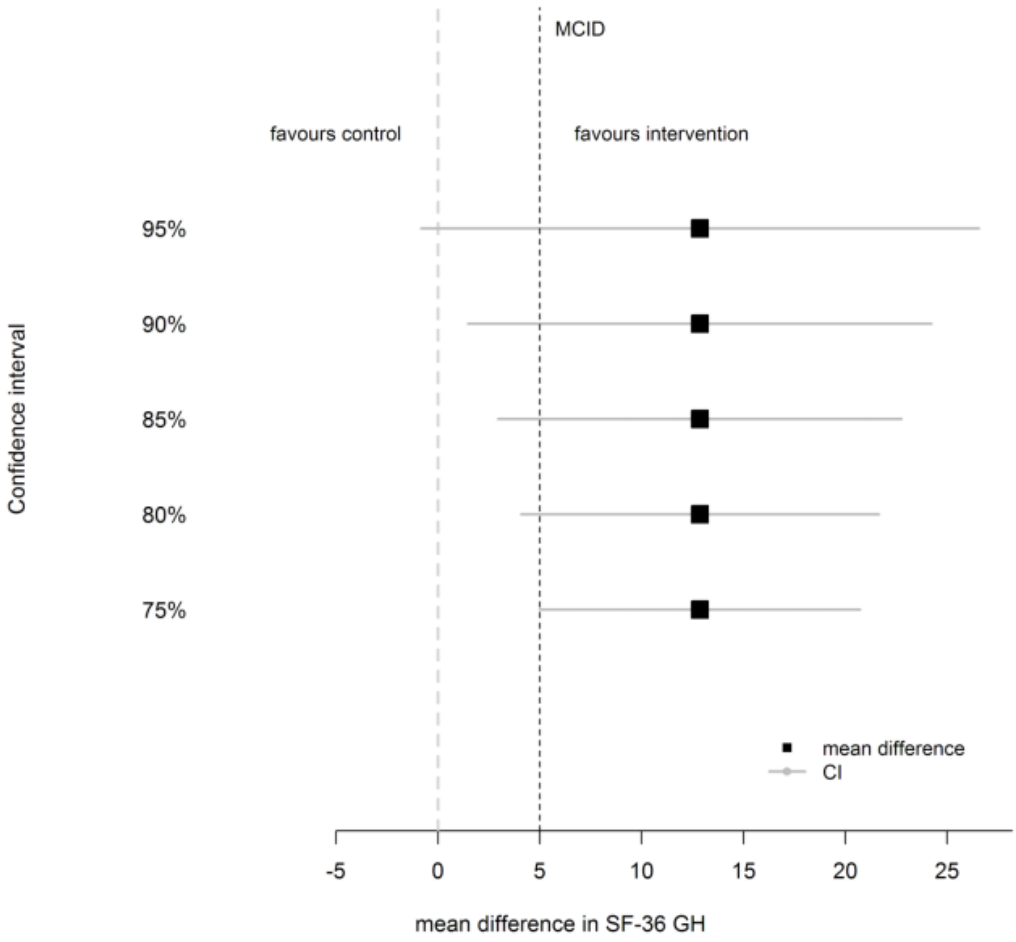
Table 1.4 shows the results of the pilot trial, the intervention group (n=17) had a mean GH score of 68.0 (SD = 17.6) and the control group had a mean score of 55.1 (SD = 19.8) giving a treatment difference of 12.8 in favour of the intervention.

Table 1.4: Mean SF-36 General Health Dimension Score for the Intervention and Control Groups

Intervention (n = 17)	Control (n = 14)	Treatment Difference (95% CI)
68.0 (SD = 17.6)	55.1 (SD = 19.8)	12.8 (-0.8 to 26.6)

Figure 1.6 demonstrates how plotting a variety of CIs alongside each other can help to display the strength of preliminary evidence, the size and direction of the treatment effect. In this example the MCID was assumed to be 5 points difference in the SF-36 GH dimension scores three months post randomisation.

Figure 1.6: Displaying a Range of Confidence Intervals



Source: (Lee et al., 2014)

The 95% CI includes both zero and the MCID giving inconclusive evidence. The 80% and 90% CIs both exclude zero and contain the MCID therefore, at these levels there is evidence of a treatment effect. This style of figure would be useful for decision making at the end of the pilot trial stage. The analysis based on confidence intervals would lead to the need for a different sample size justification than one based on a power calculation this will be discussed in more detail in Chapter 3.

1.9 Rationale and Aims

From the work presented in this chapter it can be seen that although randomised controlled trials are considered the gold standard for assessing the effectiveness of a novel intervention, they can be underpowered for the primary outcome measure if they fail to recruit participants. Only 55% (n = 40) of 73 publicly funded trials in the review by Sully et al. (2013) managed to recruit their original target sample size and 45% (n = 33) of trials received an extension (Sully et al., 2013).

A pilot trial can help to predict more accurately parameters required for the sample size calculation such as the variance of the outcome and the dropout rate (Section 1.7) and highlight issues early on in the trial development. Error rates have traditionally used acceptable values and the MCID can be derived from expert opinion. One of the main aims of a pilot may be to estimate the variance of the outcome measure likely to be observed in the main RCT. This could help to stop the waste of resources through either, inadequate sample size to fulfil a trials aims or over recruitment of patients to trials. This is especially important for publicly funded clinical trials which rely on public money. Effective allocation of public resources is of critical concern as well as the ethical conduct of clinical trials. While the methodology for confirmatory trials is well established (Pocock, 1983) (discussed further in Chapter 2) there is little guidance for conducting pilot trials (Stallard et al., 2005).

The lack of guidance is reflected in the results presented in Section 1.6 which show that although the average sample size of a pilot trial with a continuous outcome is 30, the spread is highly variable (IQR = 20-60). Compared to the work investigating required sample sizes for RCTs relatively little has been done to explore required sample sizes for pilot trials. Additionally from Section 1.6 it can be seen that 57% of trials in the UKCRN database used a continuous outcome measure as their primary endpoint. This thesis will focus on methdology for trials with a continuous outcome measure which reflects the majority of pilot trials.

The focus of this thesis will be to investigate the sample size requirements for publicly funded pilot trials of randomised controlled superiority trials with continuous outcome measures where the aim of the pilot primarily, is to provide an estimate of the variance for use in the main trial sample size calculation. As stated earlier, the thesis addresses the following objectives, to

- Provide background information on the area of pilot trials, including definitions, current sample sizes and analysis methods (Chapter 1, 3 and 6)
- Investigate how using an estimate of the variance from a pilot trial to plan a main trial affects the power and sample size of the main trial (Chapter 4 and 7)
- Explore methods of setting a sample size for pilot trials (external and internal) which aim to minimise the overall trial sample size (Chapter 4 and 7)
- Examine how the relative cost of the external pilot versus the main trial affects the sample sizes of the two trials to minimise the overall trial cost (Chapter 5)

The importance of these issues and how they will be addressed is discussed further in the literature reviews presented in Chapters 2, 3 and 6.

1.10 Outline of Thesis

This chapter began by covering some of the background information around pilot trials and clinical trial design and conduct. Developments in the sample size justifications for pilot trials are presented and the effects of pilot trials on main trial power and sample size for superiority trials with a Normally distributed continuous outcome measure are discussed in the subsequent chapters.

Chapter 2 examines the current methods for calculating the required sample size for the main trial. The methods described in this chapter are used throughout the thesis to calculate the required sample size for the main trial. The effect of using a pilot trial to estimate the parameters in the sample size calculation is also discussed in this chapter and investigated further throughout the thesis.

Chapter 3 discusses the literature on how to set the sample size for external pilot trials and introduces how the main trial sample size depends on the pilot trial sample size. The review also identifies the problems with the current methodology and states the intended scheme of work for the external pilot trial methodology chapters, Chapters 4 and 5.

Chapter 4 contains methodological work and results surrounding the setting of sample sizes for external pilot trials. To begin with the discussion is around how using an estimate of the variance from an external pilot trial affects the power and sample size of the main trial; before moving on to investigating how to choose an 'optimal' pilot trial sample size to minimise the overall size of the trial, i.e. the pilot and the main study together. This chapter identifies pilot trial sample sizes which minimise the overall sample size of the pilot and main trial together, when allowing for the imprecision of predicting the variance from a pilot trial.

Chapter 5 examines the effect on the optimal pilot study sample size if we allow for the imbalance in the costs between the pilot and the main trial and look to minimise costs instead of number of patients. Extending the work presented in Chapter 4 this chapter identifies pilot trial sample sizes which minimise the overall financial cost of the pilot and main trial together, when allowing for the imprecision of predicting the variance from a pilot trial.

Chapter 6 reviews the literature surrounding internal pilot trials. A definition of internal pilot trial is given as well as reasons why an internal pilot trial might be conducted. The review then focuses on the required size of an internal pilot trial, the proportion of the trial, which is required to be included in the internal pilot trial and the various methods for conducting a sample size re-estimation at the interim. This review introduces concepts which are investigated further in Chapter 7.

Chapter 7 addresses some of the issues surrounding the conduct of an internal pilot trial. The chapter investigates the effect conducting an internal pilot trial has on the expected

power and sample size of the main trial. Furthermore, the sample size to be used for an internal pilot trial is also explored, looking to minimise the overall trial sample size as in previous chapters. Addtionally, this chapter looks at the effect of conducting both an external pilot trial and an internal pilot trial on the expected power and sample size of the main trial.

Chapter 8 provides a summary of the conclusions of this thesis and discusses the results, implications and limitations of this work as well as ideas for possible areas of investigation for future work.

Chapter 2

Main Trial Sample Size Calculations

2.1 Introduction

It was highlighted in Chapter 1 how a sample size calculation is an important step in the design and set up of a clinical trial. A sample size calculation should reflect how the trial will be analysed (Campbell et al., 2010). The analysis of a trial depends on the trials objective, design and endpoint. For example, sample size calculations would be different if the trial is investigating superiority, equivalence or non-inferiority or whether the trial's endpoint is binary, survival or continuous etc. (Lesaffre, 2008). Descriptions of two statistical tests commonly used as a basis for sample size calculations, their implementation and reasons for choosing between them, can be found in Appendix A. Additionally, in context with this PhD the sample size and the analysis would also be different depending on whether the trial is a pilot trial or we are actually conducting a definitive assessment (Thabane et al., 2010).

The review in this chapter describes sample size calculations for a main clinical trial. Initially introducing the ideas, terminology and approaches to hypothesis testing and issues surrounding sample size calculations, the review then goes on to focus on methods for sample size calculations when the population variance is unknown and will be based on a sample estimate. Methods for sample sizes for a main trial are described here as they are used in Chapters 4, 5 and 7 in calculating main trial sample sizes in order to enable the

investigation of the method to minimise the sample size of the overall trial, the pilot and the main trial together. The concentration in this chapter is on sample size calculations for individually randomised superiority trials with Normally distributed endpoints and independent treatment groups.

The structure of this chapter is as follows: Section 2.2 describes the method of hypothesis testing and discusses some parameters which affect the required sample size of a trial; Section 2.3 explains the concept of a probability distribution function and discusses the types which will be used in this thesis. Section 2.4 discusses sample size formulae for superiority trials with a Normally distributed endpoint; Section 2.5 discusses how we might derive estimates of the parameters required for the sample size calculation. Finally, Section 2.6 provides a summary of this chapter.

2.1.1 Aims of Chapter

This chapter aims to review sample size calculation methods for superiority randomised controlled trials with independent parallel groups and a Normally distributed continuous outcome measure. In order to do so firstly some terminology and basic statistical ideas must be covered including:

- Hypothesis testing and,
- Probability distributions.

Before outlining:

- Sample size calculation formulae and,
- How to derive parameter estimates for use in the sample size calculation.

The methods reviewed in this chapter will be used throughout the thesis to enable the calculation of main trial sample sizes.

2.2 Hypothesis Testing

A clinical trial is conducted to facilitate the investigation of a research question. The term hypothesis testing describes a set of methods, which are used to analyse the findings from clinical trials. The following sections describe the methods of hypothesis testing for a superiority trial with a Normally distributed endpoint.

2.2.1 Setting up the Hypotheses

Prior to the start of the trial based on the research question we set up the hypotheses that we wish to investigate: the null hypothesis and the alternative hypothesis. The null hypothesis is usually what we wish to disprove i.e. that the mean difference between the treatment effects is zero. The alternative hypothesis is usually what we are hoping to show in our trial i.e. that the difference is not equal to zero. That is,

$$H_0: \mu = \mu_1 - \mu_2 = 0 \quad and \quad H_1: \mu = \mu_1 - \mu_2 \neq 0,$$

where μ_1 is the population mean outcome in group 1 and μ_2 is the population mean outcome in group 2 and hence μ is the population treatment effect (this could also be extended for cases where there are more than two groups). In fact we hope that the difference will be some minimum clinically relevant difference, or greater (Julious, 2009). A trial set up with a null and alternative hypothesis of this form is called a superiority trial (Gonzalez et al., 2009).

If a direction of the treatment effect is specified in the alternative hypothesis, then the test is termed one-tailed. More commonly, no direction of effect is specified in the alternative hypothesis this test would be what we call two-tailed (Altman, 1990).

2.2.2 Type I and Type II Errors

The null and alternative hypotheses relate to the population values of the mean treatment effects. However, the data we collect is only a sample from this larger whole population of interest. If we were to collect data from another sample of patients the results may differ. We therefore may by chance falsely see a difference greater than zero between the groups in our sample even if the actual difference in the whole population is zero, conversely we also may see no difference between treatments in the sample when one truly exists in the population.

When analysing the results of a trial there are two errors, which may occur, as shown in Table 2.1. You may fail to reject the null hypothesis when it is in fact false or you may reject the null hypothesis even though it is in fact true.

Table 2.1: Table to Illustrate the Idea of Type I and Type II Errors

		The Truth	
		H_0 **true**	H_1 **true**
Decision Made	**Fail to Reject** H_0	✓	Error II
	Reject H_0	Error I	✓ (Power)

Source: (Julious, 2009)

Rejecting the null hypothesis even though it is true is what we call a Type I error. The probability of a Type I error is denoted by α. Failing to reject the null hypothesis when it is false is what we call a Type II error. This is denoted by, β. This can be written as,

$$\alpha = \mathbb{P}(reject\ H_0 |\ H_0\ true) \quad and \quad \beta = \mathbb{P}(fail\ to\ reject\ H_0 |\ H_1\ true).$$

The consequence of a Type I error in a definitive trial is that a new treatment may enter the market or have results published, after having been falsely declared superior. (Julious

et al., 2010) The result of a Type II error is the failure to identify an effective treatment (Chow et al., 2003)

In reality, instead of using the value of the probability of a Type II error, β we often refer to the value, $(1-\beta)$, which is called the power. The power is the probability of rejecting H_0 when it is in fact false. Hence, in a superiority trial investigating whether one treatment is better than the other, this equates to the probability of finding a difference when one truly exists,

$$\mathbb{P}(reject\ H_0 |\ H_1\ true) = 1-\beta\ . \qquad (2.1)$$

The investigator for a trial can select the values of α and β, which they are willing to accept. It is generally desirable to keep the risk of a Type I error low at a level no higher than 5% (Wittes, 2002). The risk of a Type II error is usually allowed to be greater than that of a Type I error, set at between 0.1 and 0.2 or a power of between 80 - 90% (Julious et al., 2010) hence, a high chance of detecting a worthwhile effect if it exists (Altman, 1990).

Due to the risk of errors our data must show a difference some amount from zero for us to conclude there is a difference between the groups in the population (i.e. reject the null hypothesis). The difference that is required is defined by the type of test and the significance level or the Type I error rate for the main trial. The Type I and Type II errors are important as they will influence the size of the main trial and as a consequence, as described in Chapters 4, 5 and 7, the size of the pilot trial.

2.2.3 The P-value

If the results from our sample (or results more extreme) have a probability of less than the significance level of being seen, if the null hypothesis is true and the underlying population difference is really zero, we say that the results are statistically significant and that there is sufficient evidence to reject the null hypothesis at that significance level. This probability, the probability of seeing our results or results more extreme if the null hypothesis is true is what we call the P-value (Swinscow and Campbell, 2002).

The smaller the P-value the stronger the evidence against the null hypothesis (Kirkwood and Sterne, 2003). The observed results can be summarised by a value called the test statistic (to be described further in Section 2.2.4). The conventional cut off for the test statistic to be unlikely to be from a distribution under the null hypothesis is usually the 5% level. That is if the P-value is less than 0.05 we reject the null hypothesis that there is no difference. If the P-value is greater than 0.05 we conclude that the difference could have arisen by chance and therefore we do not reject the null hypothesis (Swinscow and Campbell, 2002). Hypothesis tests allow us to compute this P-value.

P-values are not always required in a pilot trial though they are frequently reported (Arain et al., 2010). Chapter 3 discusses the issues of using P-values in pilot trials, the situations when they should or should not be used and the options, which are available for their interpretation.

2.2.4 Test Statistics

To calculate the P-value first we calculate what is called a test statistic. When comparing two independent treatment groups the test statistic is calculated from (Altman, 1990),

$$\text{test statistic} = \frac{\bar{x}_A - \bar{x}_B}{\text{SE}(\bar{x}_A - \bar{x}_B)}, \qquad (2.2)$$

where $\bar{x}_A$ is the sample mean for treatment group A, $\bar{x}_B$ is the sample mean for treatment group B and $\mathrm{SE}(\bar{x}_A - \bar{x}_B)$ is the standard error of the difference between the sample means. The standard error of the difference between two means is calculated from the following formula,

$$SE(\bar{x}_A - \bar{x}_B) = sp\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}, \tag{2.3}$$

where sp is the pooled standard deviation estimate between the two treatment groups (defined below), $\bar{x}_A$ is the mean treatment difference in group A, $\bar{x}_B$ is the mean treatment difference in group B, n_A is the number of subjects in group A and n_B is the number of subjects in group B. The pooled standard deviation estimate is calculated from,

$$sp = \frac{(n_A - 1){s_A}^2 + (n_B - 1){s_B}^2}{(n_A + n_B - 2)}, \tag{2.4}$$

where ${s_A}^2$is the variance estimate from group A and ${s_B}^2$ is the variance estimate from group B.

The value of the test statistic is compared to the appropriate distribution table (to be discussed in Section 2.3) to find the probability of this test statistic occurring under the null distribution i.e. when the treatment difference is equal to zero, the P-value (Swinscow and Campbell, 2002).

In the context of this PhD it is the estimation of sp^2 which is the main objective of a pilot trial. The imprecision in this estimate is quantified by the degrees of freedom with which it is estimated, later it will be described how this imprecision can be accounted for when designing a main trial. In Chapters 4, 5 and 7 this approach will be used to investigate the required sample size of a pilot trial when the main trial will account for the imprecision in the variance estimate from the pilot trial.

2.2.5 Confidence Intervals

A confidence interval is a range of plausible values in which the true value of the mean difference is likely to lie. Using a 95% confidence interval would mean that under repeated sampling 95% of the confidence intervals would contain the unknown population parameter value (Swinscow and Campbell, 2002). A confidence interval should be presented alongside the result of a hypothesis test (the P-value), to aid in the interpretation of the size and direction of the treatment effect and to allow comparison to the MCID (du Prel et al., 2009). Although the 95% level is the usual confidence level used when calculating confidence intervals it is possible to calculate them using other confidence levels as highlighted (in work undertaken while doing this PhD) by Lee et al. (2014) and presented in Chapter 1.

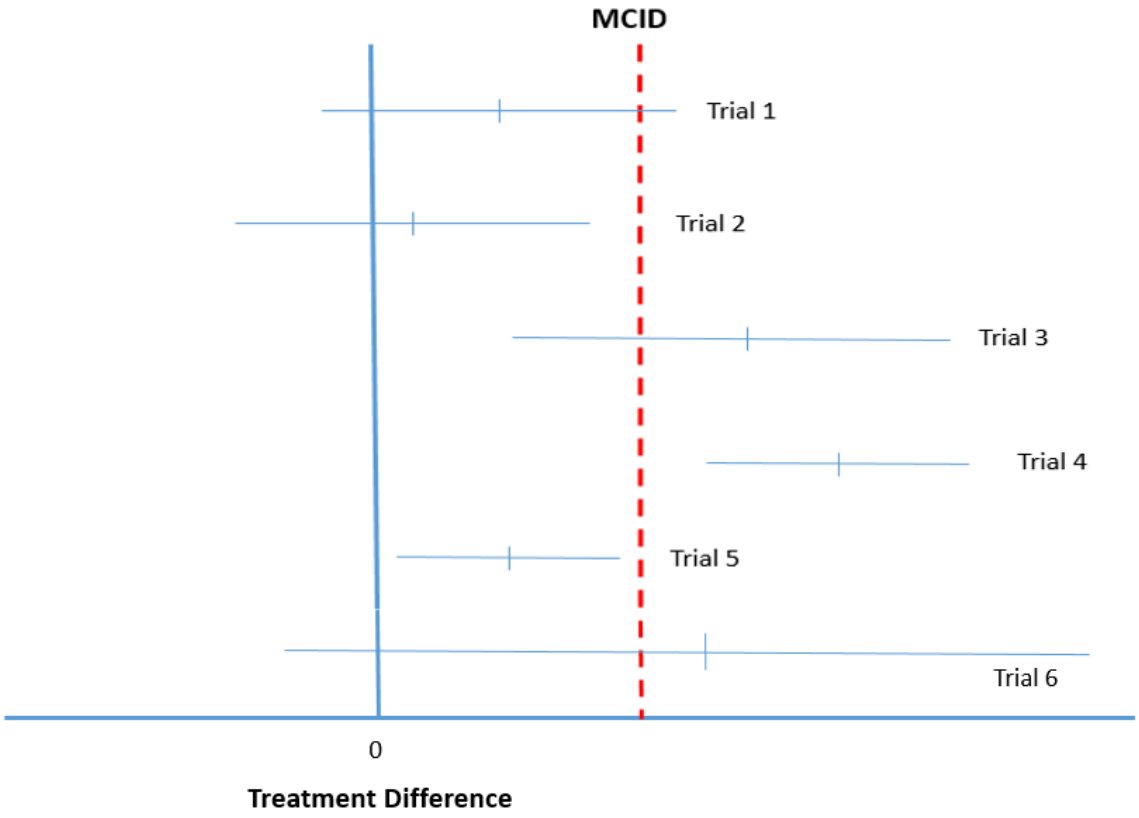
2.2.6 Statistical Significance versus Clinical Relevance

Rejecting the null hypothesis and concluding that there is a difference between the two groups does not mean that the difference seen between the groups should be considered relevant clinically. Rejecting the null hypothesis in a trial means you have shown the difference to be statistically significantly different from zero; however, the actual difference may not be different enough to be clinically relevant. It may be that one or many of the observed values of the parameters which affect the required sample size differ from the predicted values in a way which would reduce the required sample size and therefore the trial may have more power than originally planned. Alternatively, the trial could have recruited more subjects than the target sample size either through over-recruitment or through dropout being less than predicted.

For a treatment effect to be considered clinically relevant the effect should be greater than or equal to the MCID specified in the trial protocol (Stratford, 2010). This consideration of the MCID as well as the value of no difference is similar to the analysis

method using the graph in Figure 1.6 in Chapter 1. Figure 2.1 shows an example of using CIs to assess the statistical and clinical relevance of a treatment effect. For Trial 1 the CI crosses zero which indicates no effect and the dotted line which represents a treatment effect of the MCID, therefore for Trial 1 we would conclude that the result is not statistically significant but it is potentially clinically relevant. For Trial 2 the CI crosses the no effect line and does not cross the MCID line therefore, we would conclude that the treatment effect is not statistically significant or clinical relevant. For Trial 3 the line is above the no effect line and crosses the MCID line, therefore we would conclude that the treatment effect is statistically significant and potentially clinically relevant. For Trial 4 the lies wholly above the line of no effect and the MCID line, this implies that the treatment effect is both statistically significant and clinically relevant. Trial 5 shows the possibility of having a confidence interval, which lies between the line of no difference and the MCID so that the results of the trial are considered statistically significant but not clinically relevant. By sufficiently powering a trial we hope to avoid situations like Trial 6 where the difference between the two groups is estimated to be larger than the MCID however, we do not have enough data to show that this difference is statistically significant.

Figure 2.1: Using Confidence Intervals to Assess Clinical Relevance



Adapted from (Stratford, 2010)

2.3 Probability Distributions

If we have a set of events, which are mutually exclusive and contain all possible events, the sum of their probabilities is one. The set of these probabilities make up a probability distribution called the probability density function (pdf) sometimes denoted $f(y)$, where Y is a random variable and its realisation is denoted y. These distributions tend to follow patterns or recognisable distributions (Bland, 2000). Some of the more common distributions for continuous data are discussed below.

The density is a measure of the probability for a given value; however, due to the infinite number of possible values the probability of any single value will be zero. Instead we can calculate the probability of an interval in the pdf. The probability of an interval is the area under the corresponding part of the pdf (Dalgaard, 2008). Consider for example, a random variable Y, the probability of being in the interval (a, b) would be found from,

$$P(a \leq Y \leq b) = \int_a^b f(y)dy. \tag{2.5}$$

For the random variable Y the probability of being in the interval (-∞, b) or being b or less, is called the cumulative density function (cdf) sometimes written $F(y)$ (Dalgaard, 2008),

$$P(Y \leq b) = F(y) = \int_{-\infty}^b f(y)dy. \tag{2.6}$$

2.3.1 The Normal Distribution

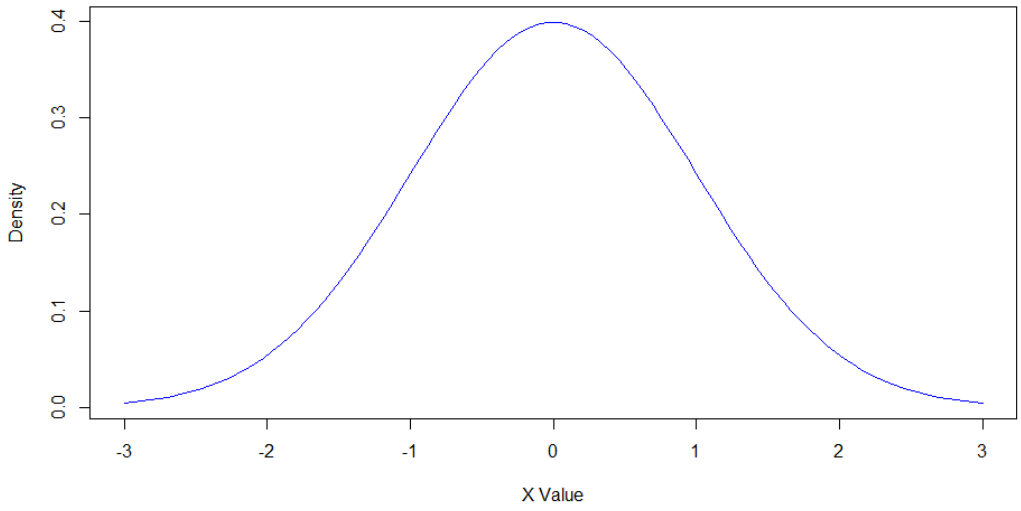
The Normal distribution is used to model continuous data that have a symmetric distribution (Dobson, 2001). Many variables can be described approximately by the Normal distribution; including many health-related measures for example, height, weight and blood pressure (Swinscow and Campbell, 2002).

This distribution is relevant to this PhD as it appears in sample size calculations for superiority trials where the data are Normally distributed, as such the details given here are used in the methodological work presented in Chapters 4, 5 and 7.

The Normal distribution can be described by two parameters: the mean, μ and the variance, $\sigma^2$and is often written as N(μ, σ^2). Figure 2.2 shows a diagram of an example of a Normal distribution. The shape of it is often described as bell-shaped as it peaks in the

middle around the mean and the spread of the distribution depends on the variance (Kirkwood and Sterne, 2003).

Figure 2.2: The Normal Distribution Probability Density Function



The probability density function of the Normal distribution with mean μ and variance σ^2 is given by,

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right). \qquad (2.7)$$

The pdf of the Normal distribution gives the height of the curve, y is any value on the horizontal axis, $\exp()$ is the exponential function and π is the mathematical constant (approximately 3.14159) (Kirkwood and Sterne, 2003). The cdf of the Normal distribution is commonly denoted as $\Phi(y)$.

The Standard Normal distribution has a mean equal to zero and a standard deviation equal to one; it has been tabulated in terms of the cdf of the distribution. The one-sided P percentage point of the distribution is the value Y such that there is a probability P% of an

observation from the distribution being greater than or equal to Y i.e. one minus the cdf of the distribution up to the value Y (this idea is used to calculate a one-sided P-value for a Z-test). The two-sided P percentage point is the value of Y such that there is a probability P% of an observation being greater than or equal to Y or less than or equal to –Y (this idea would be used to calculate a two-sided P-value for a Z-test) (Bland, 2000). The details of the Z-test are presented in Appendix A.

Changing the mean of a Normal distribution would move the distribution positively or negatively along the x-axis. Changing the variance of the Normal distribution would alter the width of the distribution; if we decrease the variance the distribution would become taller and narrower, whereas if we increase the variance the distribution would become shallower and wider (Kirkwood and Sterne, 2003).

We can change any Normal distribution to the Standard Normal distribution by subtracting the mean of the distribution from each observation and dividing by the standard deviation. For example,

$$Z = \frac{y - \mu}{\sigma}, \tag{2.8}$$

where y is the original data with mean μ and standard deviation σ, this formula gives the corresponding Z-score (position on the x-axis if that data point had arose from the Standard Normal distribution). This idea is used in hypothesis testing, discussed in Section 2.4 where our results are transformed to the Z-score so that the P-value may be calculated (Kirkwood and Sterne, 2003).

An important property of the Normal distribution is that a range of plus or minus one standard deviation from the mean will include approximately 68% of the observations, a range plus or minus two standard deviations from the mean will include about 95% of the observations additionally, adding and subtracting three standard deviations from the mean will provide a range which includes around 99.7% of the observations (Crawley,

2015). This property is essential to the theory behind confidence intervals, as it can be shown that the sampling distribution of a mean is Normally distributed (Daly and Bourke, 2008).

2.3.2 The Chi-squared Distribution

The central chi-squared distribution is formed by the sum of the squares of n independent random variables which all follow a Standard Normal distribution. If,

$$Z_i \sim N(0,1),$$

then,

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)\,, \tag{2.9}$$

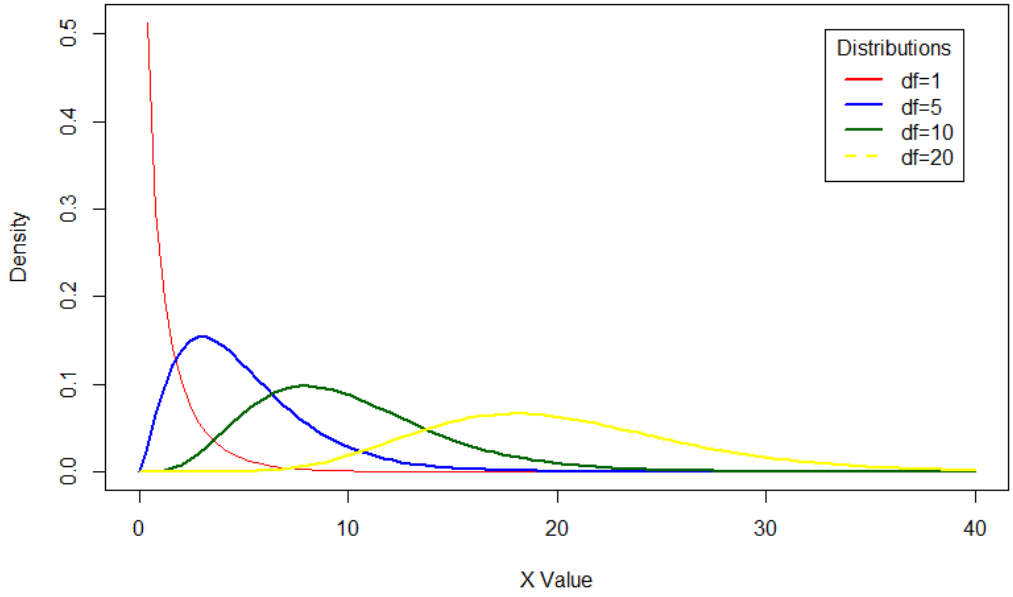
where $\chi^2(n)$ denoted the chi-squared distribution with n degrees of freedom (Dobson, 2001). Degrees of freedom are the number of independent pieces of information we have. In general this might be the sample size minus the number of constraints in a calculation, which may be the parameters that have to be estimated (Dalgaard, 2008). For example the sample variance for a single arm has $n-1$ degrees of freedom because we have to calculate the sample mean in order to estimate it. Once we have calculated the sample mean we would only need to know $n-1$ of the data points to calculate the final value, therefore we say there are $n-1$ degrees of freedom. The sample variance for a two-arm trial would have $n-2$ degrees of freedom, as we would need to calculate the sample mean for both groups to estimate the variance.

The pdf of the chi-squared distribution is given by,

$$f(y) = \frac{y^{k/2-1}\, exp(-y/2)}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)}, \tag{2.10}$$

where $\Gamma\left(\frac{k}{2}\right) = \int_0^\infty t^{\frac{k}{2}-1}\exp(-t)\ dt$, the Gamma function (Abramowitz and Stegun, 1965), y is a random variable and k the number of degrees of freedom. The shape of the chi-squared distribution varies depending on the number of degrees of freedom; this is displayed in Figure 2.3.

Figure 2.3: The Chi-squared Distribution Probability Density Function with Varying Degrees of Freedom



As n and hence the number of degrees of freedom increases the distribution tends to a Normal distribution, from the central limit theorem (Bland, 2000).

If Y_i, … , Y_n are independent random variables where,

$$Y_i \sim N(\mu_i, \sigma_i),$$

then,

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n, \lambda), \qquad (2.11)$$

because each of the variables $Z_i = \frac{y_i - \mu_i}{\sigma_i} \sim N(0,1)$. The distribution of the sum of the Y_i's, where $Y_i = Z_i + \mu_i$ is called the non-central chi-squared distribution, denoted by $\chi^2(n, \lambda)$ with n degrees of freedom and non-centrality parameter $\lambda = \sum {\mu_i}^2$ (Bland, 2000).

It can be shown that the sampling distribution of the sample variance follows a chi-squared distribution on, $k = (n - 1)$ degrees of freedom where n is the sample size (Hiorns, 1971). Therefore a one-sided upper confidence limit for the variance s^2 can be found from,

$$s_{UCL}^2 = \left[\frac{k}{\chi_{1-X,k}^2}\right] s^2 , \qquad (2.12)$$

where s^2 is the (pooled) sample variance estimate with k degrees of freedom and $\chi_{1-X,k}^2$ denotes the $1 - X$ percentile of the chi-squared distribution with k degrees of freedom (Kieser and Wassmer, 1996). This one-sided confidence interval will be used later in this chapter to define a method which recognises the imprecision involved in estimating the variance and is also used throughout the calculations presented in Chapter 4, 5 and 7.

2.3.3 The t-distribution

The t-distribution is the ratio of two independent random variables where the numerator follows a standard Normal distribution and the denominator is the square root of a central chi-squared distribution with k degrees of freedom. The t-distribution is also said to have k degrees of freedom. Therefore,

$$T = \frac{Z}{\sqrt{Y^2/k}} \sim t(k),$$

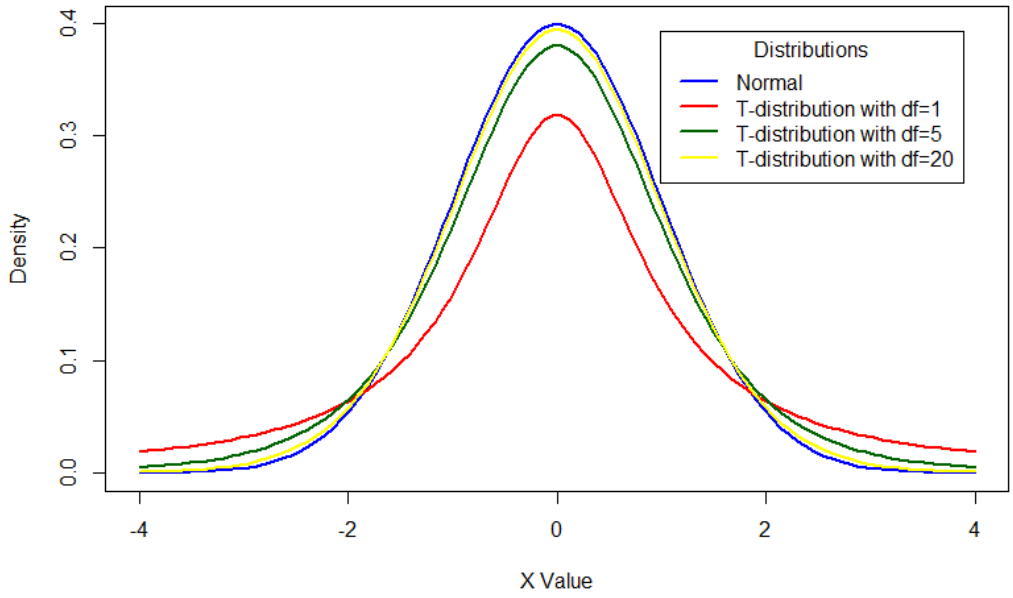
when $Z \sim N(0,1)$ and $Y^2 \sim \chi^2(k)$ and they are independent (Dobson, 2001). The pdf of the t-distribution is given by,

$$f(y) = \frac{\Gamma\left[\frac{1}{2}(k+1)\right]}{\sqrt{k\pi}\,\Gamma\left(\frac{1}{2}k\right)\left(1+\frac{y^2}{k}\right)^{(k+1)/2}}, \tag{2.13}$$

where k is the number of degrees of freedom and Γ is the gamma function.

The shape of the t-distribution is also dependent on the number of degrees of freedom and as this increases the t-distribution tends to the Normal distribution. The shape of the t-distribution can be seen (Figure 2.4) to be very similar to the Normal distribution but it is slightly thicker at the tails with a shallower peak.

Figure 2.4: The T-distribution Probability Density Function with Varying Degrees of Freedom compared to the Normal Distribution



The sampling distribution of the mean is Normal and the sampling distribution of the variance follows a chi-squared distribution, therefore the ratio of a sample mean and its standard error will follow a t-distribution (Bland, 2000). This result is used in the method presented as the non-central t-distribution later in this chapter, which is used throughout the methodological work presented in Chapters 4, 5 and 7.

The non-central t-distribution is a t-distribution that is not centred around zero which can be expressed as,

$$T = \frac{Z + \lambda}{\sqrt{Y^2/k}} \sim t(k, \lambda),$$

where λ is the non-centrality parameter. In figure 2.10 the t-distributions can be seen to all be centred around zero. This is achieves by setting the non-centrality parameter, λ in

the equation above to be zero. Therefore, the non-central t-distribution could be thought of as the generalised version of the standard central t-distribution.

From Figure 2.5 it can be seen that by changing the non-centrality parameter we can change how much the distribution is shifted along the horizontal axis. Changing the non-centrality parameter to a positive number moves the distribution to the right along the x-axis. Changing the non-centrality parameter to be a negative number moves the distribution down the x-axis. These are also plotted alongside the corresponding non-central Normal distribution to show how the shape of the t-distribution varies from that of the Normal distribution.

From Figure 2.6 it can be seen how changing the degrees of freedom for the non-central t-distribution changes the shape of the distribution. When the distribution is non-central and the number of degrees of freedom is small, the distribution can become skewed. As the number of degrees of freedom increases the distribution becomes more symmetric and is centred around the non-centrality parameter. The corresponding non-central Normal distribution is also plotted to show that as the number of degrees of freedom increases the t-distribution tends to the Normal distribution.

The Type I error rate can be estimated by rearranging the sample size formula assuming the null hypothesis is true therefore assuming no difference between the treatment difference. The Type II error rate can be estimated assuming the alternative hypothesis is true, therefore that the treatment difference is equal to the minimum clinically important difference. Generally, this is done by using the central Normal distribution (mean of zero) and non-central Normal distribution to calculate the Type I and II errors respectively. However, when there are small numbers involved an approximation to the Normal distribution is used, the t-distribution. When this is the case, the estimation of the Type I error and the Type II error can be done by using the central and the non-central t-distribution respectively.

Figure 2.5: The Non-Central t-distribution Probability Density Function with Constant Degrees of Freedom (k=5) with Varying Non-Centrality Parameters (-5, 0, 5) compared to the Normal Distribution

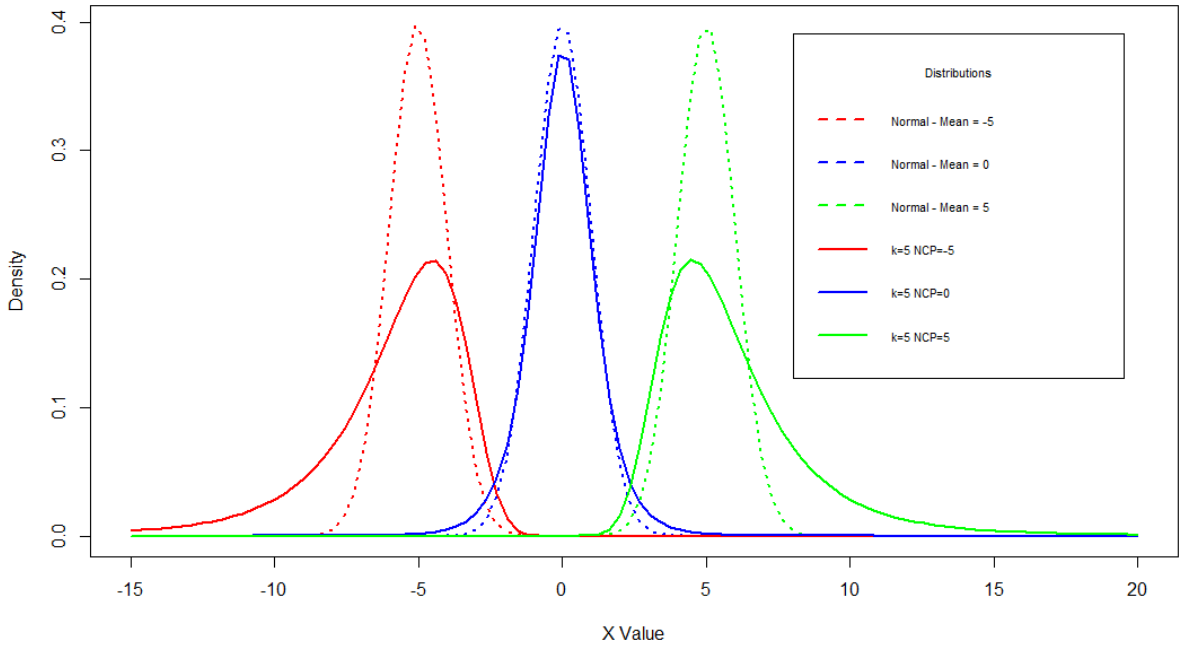
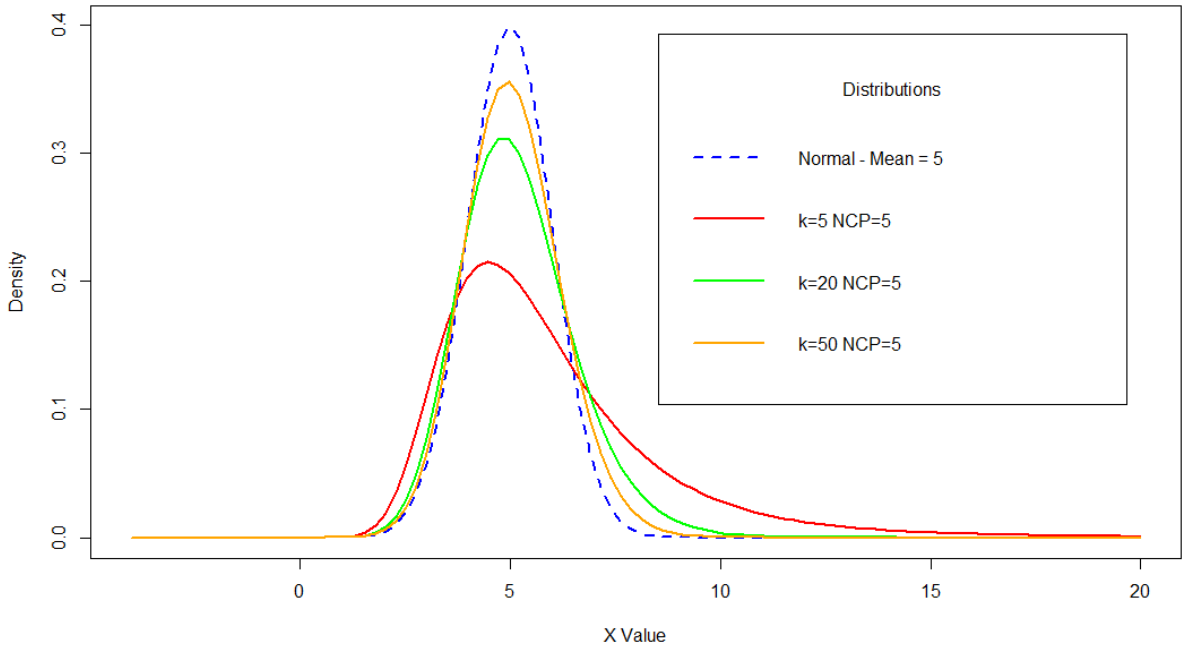


Figure 2.6: The Non-Central t-distribution Probability Density Function with Varying Degrees of Freedom (k= 5, 20 and 50) compared to the Normal Distribution



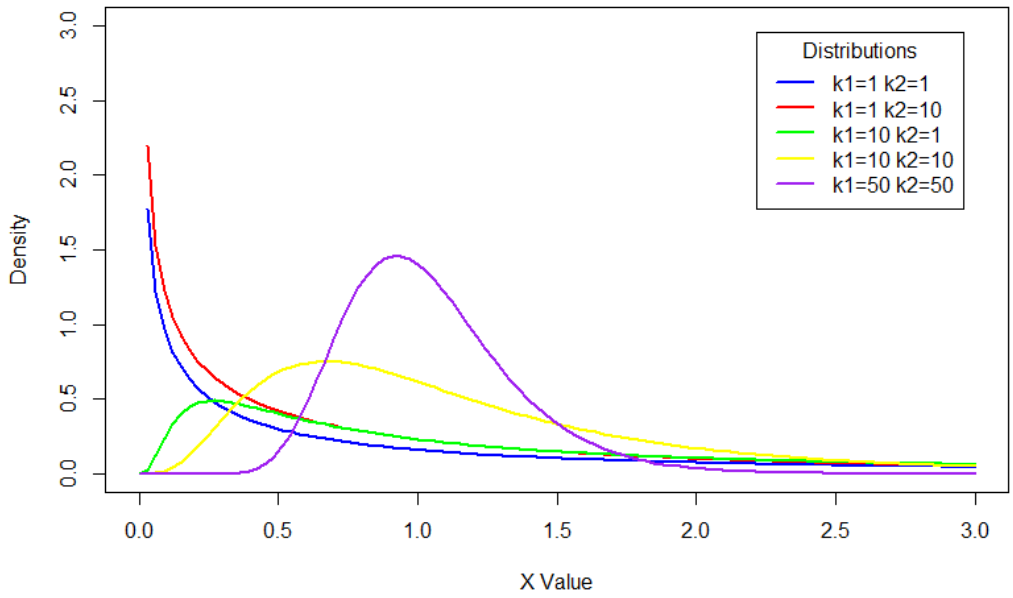
2.3.4 The F-distribution

The central F-distribution is the ratio of two independent central chi-squared random variables each divided by its degrees of freedom k_1, k_2. Therefore,

$$F = \frac{Y_1^2}{k_1} \Big/ \frac{Y_2^2}{k_2} \sim F(k_1, k_2),$$

where $Y_1^2 \sim \chi^2(k_1)$, $Y_2^2 \sim \chi^2(k_2)$ and these are independent (Dobson, 2001). The shaped of the F-distribution is affected by both k_1 and k_2. Some example distributions can be seen in Figure 2.5.

Figure 2.7: The F-distribution Probability Density Function with Varying Degrees of Freedom



The non-central F-distribution is defined as an F-distribution where the numerator follows a non-central chi-squared distribution (Dobson, 2001). This distribution is used in the calculations presented as part of the work in Chapter 7 on internal pilot trials, where the ratio of the sample variances (both of which are chi-squared distributed) from the external pilot and the internal pilot is considered.

2.4 Sample Size Formulae

The following section describes the methods available for calculating the required sample size for a definitive trial, where the analysis will be a hypothesis test on the mean difference between the two independent treatment arms, and the outcome is continuous and Normally distributed for a superiority trial.

A sample size formula tries to strike a balance between the trial sample size being too large or too small by calculating the minimum number of participants needed to ensure

the required power and control Type I error rate; for a specified treatment difference and variance. The factors that affect the required sample size of a superiority trial are: The Type I and Type II error levels, the allocation ratio of patients between treatment groups, the variance of the outcome measure and the MCID.

2.4.1 Z-test

By using the power statement set out in Equation 2.1 the following formula can be derived (full derivation can be found in Ch. 6 of *'Statistical Methods'* (Snedecor and Cochran, 1989)) for calculating the required sample size.

$$n = \frac{(r+1)}{r} \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma^2}{d^2}, \tag{2.14}$$

where, n is the sample size per treatment arm, $Z_{1-\beta}$ corresponds to the standard Normal distribution Z-score of (1 – β) and $Z_{1-\alpha/2}$ is the Standard Normal Z-score of (1 - α/2), σ^2 is the variance estimate, d is the minimum clinically important difference (MCID) and r represents the allocation ratio of participants between the treatment and placebo group of r: 1. The sample size n is the number of participants in the placebo group, the number of participants in the treatment group is given by nr, where r is the allocation ratio. The value of $Z_{1-\alpha/2}$ can be found from using statistical software or from statistical tables. An example of a table of the Normal distribution can be found in Appendix B. If for example, $\alpha = 0.05$ (two-sided) we wish to calculate the value for $Z_{0.975}$, we look up 0.975 in the probability column this gives a Z-score of 1.96, or for a power of 90% ($\beta = 0.1$) $Z_{0.9} = 1.28$. For a superiority trial with equal allocation between treatment groups the sample size formula would be,

$$n = \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma^2}{d^2}. \tag{2.15}$$

Using a two-sided significance level of 5% and a required power of 90% this formula approximately reduces to,

$$n = \frac{2(1.28 + 1.96)^2 \, \sigma^2}{d^2}$$

$$n = \frac{21\sigma^2}{d^2}.$$

(Julious et al., 2010)

Throughout this thesis r is assumed to be 1 and the number of treatment arms is assumed to be 2. Therefore going forward Equation 2.15 will be used when required and r will no longer be included in the sample size equations explicitly. Operationally the Z-test is used as a large sample approximation to a t-test, however, strictly the Z-test should only be used if the population variance is known. In practice the variance used is an estimate derived from previous work or, in the context of this PhD, a pilot trial.

2.4.2 t-test

The previously presented method based on a Z-test assumes that the variance of the outcome measure is known at the design and analysis stage. However, in reality the variance used in the analysis of a trial is an estimate, s^2 from the sample data, of the population variance, σ^2. Some modification is needed to the analysis and the sample size calculation to account for this (Wittes, 2002). A t-test will be used instead of a Z-test to allow for this estimation. The sample size calculation should always reflect the type of analysis, which will be carried out, therefore because we are using a sample variance in the test statistic rather than the known population variance. The following sample size formula assumes the analysis will be based on a t-test,

$$n \geq \frac{2\sigma^2 \left(Z_{1-\beta} + t_{1-\alpha/2,2n-2}\right)^2}{d^2}, \tag{2.16}$$

which can be re-arranged to show that,

$$\begin{aligned} \frac{nd^2}{2\sigma^2} &\geq \left(Z_{1-\beta} + t_{1-\alpha/2,2n-2}\right)^2 \\ \sqrt{\frac{nd^2}{2\sigma^2}} &\geq Z_{1-\beta} + t_{1-\alpha/2,2n-2} \\ Z_{1-\beta} &\leq \sqrt{\frac{nd^2}{2\sigma^2}} - t_{1-\alpha/2,2n-2} \\ 1-\beta &\leq \phi\left(\sqrt{\frac{nd^2}{2\sigma^2}} - t_{1-\alpha/2,2n-2}\right) \end{aligned} \tag{2.17}$$

Here the power is estimated from a cumulative Normal distribution. However, replacing σ^2 with s^2 Equation 2.17 becomes,

$$1-\beta = P\left(\sqrt{\frac{nd^2}{2s^2}} - t_{1-\alpha/2,2n-2}\right), \tag{2.18}$$

where P(.) denotes a probability which can be shown to follow a t-distribution (Julious, 2009). Equation 2.17 represents an approximation to Equation 2.18 for when the variance is assumed to be known at the start of the trial but the analysis will be based on an estimate of the variance from the data at the end of the study. As n appears on both sides of the Equation 2.16 it must be solved using iteration. A good starting point is the sample size given in Equation 2.15 with s^2 replacing σ^2 (Julious, 2009).

The test statistic at the end of the trial is calculated using the method described in Section 2.2.4, however the value is now compared to the tables of the t-distribution along with the degrees of freedom (df), here, $df = n_1 + n_2 - 2$, to find the P-value (Swinscow and Campbell, 2002).

2.4.4 Dropout Rate

The number of participants calculated from the sample size formula is the required number of evaluable patients that need to be remaining at the end of the trial. During the trial some people will be lost to follow up and therefore the recruitment target at the start of the trial will be the number from the sample size calculation plus extra to allow for this dropout. The recruitment target at the start of the trial to allow for patient dropouts N^* can be calculated from,

$$N^* = \frac{n}{1 - W}, \tag{2.19}$$

where n is the sample size per group from the sample size calculation formula and W is the dropout or withdrawal rate (Campbell et al., 2010). We must also remember that not all patients who are eligible will consent to be involved in the trial; some estimate of this proportion is also needed to calculate the target number of patients to approach, as discussed in Chapter 1.

2.5 Deriving Parameters for a Main Trial Sample Size Calculation

When planning a trial firstly the question must be defined by setting the scientific hypotheses we are trying to test, outlining the objective for the trial and choosing the primary endpoint(s), which will be used for the analysis. To further calculate the required sample size, the investigator needs to specify the minimum clinically important difference (MCID) between the two treatments in the trial, the Type I and Type II error rates and the variance of the outcome measure (Friede and Kieser, 2001).

Although it can be a complex procedure (Wright et al., 2012), clinical expertise can be used to specify the MCID and there are traditional values used for the Type I and Type II error levels (Section 2.2.2). The difficulty comes when trying to specify the variance (Friede and Kieser, 2001). Setting the variance estimate at an inappropriate level can have a serious effect on the power of the trial (Denne and Jennison, 1999). If the anticipated variance before the trial is greater than the true variance, then the trial will be overpowered. If the estimate is too low, then the trial will be underpowered to find the effect.

Investigators can use several different methods to try to get an accurate estimate of the true variance of the outcome measure. They may use historical data to calculate an estimate of the variance i.e. past published studies or meta-analyses (Wittes, 2002), conduct an external pilot trial prior to the trial (see Chapter 3 and 4) or conduct an internal pilot trial (see Chapters 6 and 7). The variance can be difficult to estimate, allowing for this difficulty is the focus of Section 2.5.2 and the methodological work in Chapter 4, 5 and 7. In this thesis the concentration is on conducting pilot trials to estimate the variance to be used in the main trial sample size calculation.

2.5.1 Issues with using Historical or Pilot Data to Plan the Main Trial

Estimations from historical data can be misleading. Using a large amount of papers from the literature to estimate the variance to be seen in a single study is likely to produce an underestimate of the variance (Wittes and Brittain, 1990). Published studies are likely to underestimate the variance on average, due to publication bias. Underestimated variances lead to higher probabilities of statistically significant results hence higher chance of publication. (Wittes, 2002) Those studies, which overestimate the variance, are less likely to find statistically significant results and hence are less likely to get published (Torgerson and Torgerson, 2008); therefore it is likely that the literature underestimates the variance parameter. In addition, the trials may have been conducted in a different population from the planned trial and under different trial conditions (Denne and Jennison, 1999).

A pilot trial may also differ from the main trial in ways, which affect the variance estimate. They may: include very few centres, have different inclusion/exclusion criteria, be conducted in different populations, have different primary endpoints or use a surrogate, be of shorter duration and the treatment duration may differ from that intended in the main trial (Kianifard and Islam, 2011; Wittes and Brittain, 1990).

Even if the pilot trial has the same design as the main trial, by definition pilots are small and therefore may be subject to sampling error (Denne and Jennison, 1999) which can in turn mislead the sample size calculation (Kraemer et al., 2006). A study conducted by Vickers in 2003 found that around 80% of the time an estimate of the standard deviation used in the main trial sample size calculation was an under-estimate and hence trials are usually underpowered. In the following sections, methods will be described for sample size calculations for main trials that account for the imprecision of the variance estimate.

2.5.2 Methods for Overcoming the Issues with using Pilot Data

Using previous trial results to estimate the variance introduces a type of imprecision that is not allowed for in the sample size calculations described earlier in this chapter. This section discusses methods proposed to try to deal with this problem of inaccurate estimation of the variance by adjusting the prediction of the variance from a pilot trial for use in a main trial sample size calculation.

2.5.2.1 The Upper Confidence Limit Approach

One such method will be referred to as the Upper Confidence Limit (UCL) approach. It was proposed by Browne (1995), who carried out simulations, which suggested that using an 100X% upper confidence limit for the estimated value of the variance from the pilot trial to plan the main trial would provide a sample size sufficient to achieve the required power in at least 100X% of such trials. Kieser and Wassmer (1996) later proved this method mathematically in 1996. Therefore, X is a probability, which is set at a level we are willing to accept for achieving the required power.

In order to implement the UCL approach, a variance estimate is obtained and the one-sided 100X% upper confidence limit for it, s^2_{UCL} is calculated, from the equation below as stated in Equation 2.12 in Section 2.3.2, where k is the number of degrees of freedom. This one-sided upper confidence limit can be calculated based on the chi-squared distribution as stated earlier from (Kieser and Wassmer, 1996),

$$s^2_{UCL} = \left[\frac{k}{\chi^2_{1-X,k}}\right] s^2 .$$

This UCL would then be used as the variance estimate in the sample size equation (Equation 2.16). Therefore, the power for the main trial would be given by,

$$1 - \beta = P\left(\sqrt{\frac{nd^2}{2s_{UCL}^2}} - t_{1-\alpha/2,2n-2}\right). \quad (2.20)$$

This method will be used later in Chapter 4 when investigating the relationship between pilot trial sample size and the main trial sample size when the main trial sample size is calculated using this approach. These calculations will also be compared to the sample sizes required if the main trial sample size is calculated using the alternative approach presented in the next section (Section 2.5.2.2).

Although he recommends an 80% confidence level, Browne gives no explanation or justification for this figure and so Browne's method leaves open the question as to what level should be set. Sim and Lewis (2011) set X at the 95% level however it seems sensible to question whether this level of rigour is required.

2.5.2.2 The Non-Central T-Distribution Approach

Julious and Owen (2006) proposed an alternative method for the calculation of sample size, which accounts for the fact that we are using s^2 (from a sample) instead of σ^2 (the population value) in the sample size calculation. Using previous trial results to estimate the variance introduces a type of imprecision that is not allowed for in Equation 2.16. The sample size is inflated dependent on the number of degrees of freedom k, on which the estimate of the variance is based, through Equation 2.21,

$$n \geq \frac{2s^2[tinv\left(1-\beta, k, t_{1-\alpha/2,2n-2}\right)]^2}{d^2}, \quad (2.21)$$

where, $tinv(\cdot, k, a)$ is the inverse function of the cumulative distribution function of a t-distribution with a non-centrality parameter, a on k degrees of freedom. Here k is the degrees of freedom about the estimate s^2. This method is derived from integrating not

only over the Normal distribution centred around the effect size but also over every plausible value of the variance (from a chi-squared distribution) to find the expected power. As shown by Julious and Owen (2006) this solves to be a non-central t-distribution, therefore from here on this method will be referred to as the Non-Central T-distribution (NCT) approach.

This must be solved iteratively as n appears on both sides of the inequality. A good starting point for the iterations can be found from Equation 2.22 as outlined by Julious and Owen (2006),

$$n = \frac{2s^2[tinv\left(1-\beta, k, Z_{1-\alpha/2}\right)]^2}{d^2}. \quad (2.22)$$

If the estimate of the variance is based on only a few degrees of freedom the sample size will be increased greatly, as the number of degrees of freedom for the estimate of the variance increases the effect of this method on the sample size decreases. As the degrees of freedom, k increases Equation 2.21 tends to Equation 2.16, furthermore as the number of degrees of freedom increases the sample variance will also tend to the population variance. Julious and Owen (2006) show that this happens after the degrees of freedom reaches 200. This occurs because as the sample size increases the t-distribution tends to a Normal distribution. The higher the number of degrees of freedom: the less sensitive calculations are to assumptions about the variance (Julious, 2004).

Due to the difficulty involved with specifying a MCID and a variance estimate it is common to instead specify the standardised difference, that is, the MCID divided by the standard deviation. The standardised difference is denoted by δ and is given by $\delta = {}^{d}/_{s}$, therefore, $\delta^2 = {}^{d^2}/_{s^2}$ and ${}^{s^2}/_{d^2} = {}^{1}/_{\delta^2}$. It is therefore possible to replace the d and the s in Equation 2.22 with δ giving:

$$n = \frac{2[tinv\left(1-\beta, k, Z_{1-\alpha/2}\right)]^2}{\delta^2}.$$

Cohen (1992) proposed the use of small, medium and large standardised effect sizes of 0.2, 0.5 and 0.8 respectively in order to allow comparisons of effect sizes across scales. This approach also allows the selection of an effect size when there is little information about the required difference between two treatments in a trial. This idea will be used in Chapter 4, 5 and 7 to aid in the calculation of general sample sizes.

2.6 Summary

This chapter summarises the literature on sample size calculations for superiority trials with two independent treatment arms and a Normally distributed continuous outcome measure. It initially describes the process of hypothesis testing before moving on to outline the methods of sample size calculation for main RCTs. The procedures presented will be used in Chapters 4, 5 and 7 for the calculation of sample sizes for main trials.

The required sample size for the main trial depends on the allocation ratio of patients between the treatment groups, the Type I and Type II error levels, the variance of the primary outcome measure and the MCID. Prior to carrying out the sample size calculation estimates of these parameters will be needed. Keeping the risk of errors low requires a higher number of participants than if the risk of errors was allowed to rise. A tighter control over the level of errors will require more trial participants. The most efficient allocation ratio of patients between treatment groups is 1:1, which is equal allocation. Deviating from this will increase the required sample size for the trial. The sample size required is proportional to the variance of the outcome measure. As the variability within the data increases we will require more people within the trial. If the variability is low, we will require relatively few participants. The larger the MCID the fewer participants are needed within the trial. If we reduce the treatment effect size, we are looking for the required size of the trial will increase.

Pilot trials can be used to estimate the variance for the main trial sample size calculation; the imprecision of these estimates, can impact on the accuracy of the calculation (Kraemer et al., 2006). Two methods for adjusting the sample size calculation to allow for this uncertainty were presented; the UCL method and the NCT approach. The impact of these methods will be investigated in Chapter 4.

The procedures presented in this review are for main trials and may not be applicable for pilot trials. However, they are used in further chapters (4, 5 and 7) to investigate the

impact the pilot trial sample size has on the sample size of the main trial. Chapter 4 and 5 investigates the procedures for choosing a sample size for an external pilot trial and the review in Chapter 6 looks at the sample size requirements of an internal pilot trial.

Chapter 3

Pilot Trial Sample Size Justifications

3.1 Introduction

The previous chapter outlined methods for hypothesis testing and sample size calculations for a main RCT. In Chapter 1 it was discussed how sample size methods based on hypothesis testing may not be appropriate for pilot trials and in Chapter 2 methods for sample size calculation for a main trial were described.

The sample size calculations described in Chapter 2 assume that the analysis of the trial will be based on a hypothesis test where the null hypothesis may or may not be rejected. The sample size calculations thus rely on an expression of the power of the trial or the probability of rejecting the null if it is false; however, this is not always appropriate for a pilot trial.

Extending the work highlighted in Chapters 1 and 2 this chapter discusses methods of choosing a sample size for an external pilot trial. The structure of this chapter is as follows: Section 3.2 presents reasons why the standard sample size calculation methods may not be appropriate for pilot trials; Sections 3.3 and 3.4 discuss the current methods used for choosing an external pilot trials sample size. Finally, Section 3.5 provides a summary for this chapter and describes the aims for Chapters 4 and 5, which are derived from the reviews presented here and in Chapter 2.

3.1.1 Aims

This chapter aims to:

- Identify existing methods for justifying a pilot trial sample size,
- Establish any weaknesses with the current approaches and therefore,
- Outline the areas of work for Chapters 4 and 5.

3.2 Standard Sample Size Calculations and Pilot Trials

For a main trial incurring a Type I error would result in a new treatment being falsely assessed as superior to the control treatment (Julious et al., 2010) and a Type II error would mean that an effective treatment would not be taken forward and would be assessed as inferior – or no better - than the control treatment (Schoenfeld, 1980).

As previously highlighted in Chapter 2 for a pilot trial Type I and Type II errors have different implications than within a main trial. A Type I error may lead to a large definitive trial being incorrectly run and hence could be an expensive mistake for the sponsor. However, this consequence is different from that of a Type I error in a definitive trial, a Type I error in a pilot trial has the chance to be corrected in the definitive trial so that the inferior treatment would not make it to the market (Stallard et al., 2005).

As the consequences are less severe if a Type I error is made in a pilot trial than a main trial authors have suggested that the Type I error rate could be increased (Stallard, 2011). There could also be scope to reduce power for a pilot trial compared to the formal levels used in definitive trials (Stallard, 2011); although it may seem more desirable to leave the power high to avoid losing any effective treatments. Methods for powered pilot trial sample size calculations are discussed in Section 3.3.

There are a number of arguments against doing a formal sample size calculation for a pilot trial. A main reason is that the decision whether to carry on to the definitive trial at the

end of the pilot trial is often based on more than one criterion; for example, the acceptability/ safety of the intervention, the recruitment rate estimate, the dropout rate, the feasibility of the definitive trial sample size based on the estimated variance from the pilot trial (Thabane et al., 2010).

Pilot trials have different aims and objectives to main trials as in the definitions laid out in Chapter 1. Pilot trials are not looking to prove superiority of the experimental treatment yet, they are looking to test trial procedures and processes and get estimates of parameters for the main trial sample size calculation (Lancaster et al., 2004). So the traditional sample size determination methods (as seen in Chapter 2) based on hypothesis testing seem inappropriate for pilot trials as these focus on getting the required number of people to test the superiority of one treatment over the other (Thabane et al., 2010). In addition, referring back to the requirement of the CONSORT statement and bodies such as NIHR and NRES; all studies do not necessarily need a sample size calculation but they do all need a sample size justification, therefore, other criteria have been developed which may be used to set an unpowered sample size. These are presented and discussed in Section 3.4.

3.3 Powered Calculations

Traditional sample size calculations can be inappropriate for pilot trials however, if at the end of the pilot trial the analysis will be to compare the interventions through a hypothesis test; then the sample size should be set based on a power calculation, in order to ensure that the required sample size to give a specified chance of seeing a difference if one exists is known. Labelling the trial a pilot should not be an excuse to run a small underpowered trial, which would have little scientific validity and would therefore be unethical (Arain et al., 2010, Halpern et al., 2002). However, we may not need to use the same error rates as are conventional in definitive trial sample size calculations as discussed in Chapter 1 Section 1.8.

Situations could arise for example, if perhaps the pilot trial is designed based on a surrogate or biomarker endpoint for the true clinical outcome. Here there could be a formal powered sample size calculation but it would not be on the primary outcome of interest, as would be used in the main trial. Estimates needed for the sample size calculation based on the surrogate endpoint could be gained from similar trials with the same endpoint or from data about the control treatment.

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) guideline E9 defines a surrogate endpoint as 'A variable that provides an indirect measurement of effect in situations where direct measurement of clinical effect is not feasible or practical' (ICH, 1998, p.35). Surrogates can be useful if the clinically relevant outcome is long in duration (e.g. the 5 year survival rate) (DeGruttola et al., 2001), invasive, uncomfortable or expensive (Prentice, 1989). Using a surrogate endpoint in a pilot trial can therefore reduce the cost, duration and sample size of the trial, compared with using the true clinical outcome (Temple, 1999). Conducting a trial based on a surrogate could mean that it is very difficult to get accurate estimates of parameters for the main trial sample size calculation, which would be based on the clinical endpoint to be used in the main trial.

For pilot trials where the analysis will be based on an hypothesis test Schoenfeld (1980) proposed maintaining the rigorous power requirement of 90%, as usual in a main trial, but letting the Type I error rate be up to 25%. He commented that a Type II error is more serious in a pilot trial as this would mean denying an efficacious treatment from the market. However, a Type I error could be corrected in the definitive main trial. Therefore, advocating reasonable levels of α and β as 0.25 and 0.1 respectively. As highlighted in Chapter 1 Section 1.8, Stallard (2011) recommends to design a pilot trial with a one-sided Type I error rate of 0.2 and a power of 40%. His design minimises the average number of patients required per successfully identified effective therapy; therefore, reflecting the costs involved in conducting a series of pilot studies and definitive trials prior to a successful definitive trial. These recommendations are summarised in Table 3.1.

Table 3.1: Powered Sample Size Power and Type I Error Recommendations

Author	Power Level	Type I Error Level	Relative Sample Size
Conventional	90%	5%	1.00
Schoenfeld (1980)	90%	25%	0.56
Stallard (2011)	40%	20%	0.10

The last column in Table 3.1 gives the relative sample size of the trials designed with parameters compared to the conventional levels usually used. Therefore, a trial designed to the specifications laid out by Schoenfeld would require 56% of the number of patients required for a conventional design with 90% power. The design of Stallard would require only 10% of the patients required under the conventional design.

3.4 Unpowered Sample Size Justifications

If the purpose of the pilot is not to give a preliminary assessment of treatment effect to show proof of concept, then the sample size provided by the conventional calculations is often higher than necessary (Lancaster et al., 2004), however the sample size of the trial still needs to be justified. Unpowered sample size justifications may be carried out when a sample calculation based on formal considerations is thought to be inappropriate. For example, if the purpose is not to give an assessment of efficacy but to estimate, the treatment effect, the variance of outcome measure, accrual or attrition etc.

Although not the focus of this thesis, there have been recommendations made for pilot trial sample sizes if the outcome is a proportion and in particular around feasibility features in trials, for example, the accrual or attrition rates. Examining this question was another focus of the paper by collaborators in ScHARR Teare et al. (2014). This work found that the relative gain in precision of the variance estimate dropped below the 5%

threshold level once the sample size of the pilot trial was between 55 and 60 and dropped below the 3% level when the sample size was 100, these sample sizes are per treatment group for estimating an event rate. It would also be possible to use the precision based methods as presented in Section 3.4.1 using the CI for a proportion (Cocks and Torgerson, 2013). The recommendations discussed in this chapter for unpowered sample size justifications are summarised in Tables 3.2, 3.3 and 3.4.

3.4.1 Precision-Based Calculations

At the beginning of treatment development, estimates of treatment effect may not be available and therefore it is not possible to estimate sample size based on a treatment difference of interest as in the powered sample size section. For such studies, Julious and Patterson (2004) present a way of calculating the necessary sample size by setting the required precision for the possible clinical effect. This method is recommended by Thabane et al. (2010) in their tutorial on pilot studies for setting sample sizes required to establish feasibility. This method also complements the ideas of Lancaster et al. (2004) that the analysis of a pilot should be based on confidence interval estimation; referred to as the estimation approach.

The formula for a CI for a continuous outcome measure is shown below

$$\bar{x} \pm t_{1-\alpha/2,k} \sqrt{\frac{\sigma^2}{M}}, \qquad (3.1)$$

where, k is the number of degrees of freedom, M is the total pilot trial sample size, $\bar{x}$ is the treatment difference. In order to implement this approach, we must have a specification for the required width of the CI. We call the half width of the CI, w.

$$w = t_{1-\alpha/2,k} \sqrt{\frac{\sigma^2}{M}}, \qquad (3.2)$$

The formulas can be re-arranged to calculate the sample size required to give a certain precision as shown below (Julious and Patterson, 2004)

$$M = \frac{\sigma^2 \, t^2_{1-\alpha/2,k}}{w^2}.$$

Parker and Berman in 2003 wrote that a different approach would be more appropriate for small studies such that, instead of considering the sample size needed for a certain amount of information, consideration is given as to how much information we can get for a certain sample size. This could be achieved using the Julious and Patterson method by considering what precision could be achieved for a given sample size and deciding whether this level of precision is appropriate to meet the investigator's goals (Kianifard and Islam, 2011).

3.4.2 Flat Rules of Thumb for Selecting Pilot Trial Sample Size

Several rules of thumb have been suggested for use when setting the sample size of pilot trials for definitive clinical trials, these are summarised here and presented in Table 3.2.

Table 3.2: Flat Rules of Thumb for External Pilot Trial Sample Sizes (for a Two-Armed Trial)

Author	Type of Calculation	Inflation Method	Recommended Sample Size
Julious (2005)	Precision based		24
Browne (1995)	Precision based		30
Teare et al. (2014)	Precision based		70
Sim and Lewis (2012)	Drop in main trial power	90% UCL	60
Sim and Lewis (2012)	Drop in main trial power	95% UCL	50

Browne (1995) cites a general rule to use at least 30 subjects or greater to estimate a parameter, this parameter could be for example, the variance of the outcome measure. Julious (2005) presents three reasons for a minimum sample size of 24 (12 subjects per treatment arm); feasibility, precision around the mean (based on the width of the CI as given in Equation 3.2) and variance estimates and regulatory considerations. In anticipation of planning the definitive trial we could also carry out a sensitivity analysis for the drop in power for the future trial if the variance turned out to be at the 95th percentile from the pilot trial, calculated from, (Diem, 1962)

$$s^2(95) < \frac{df}{\chi^2_{0.95,df}} s^2 \qquad (3.3)$$

Julious (2005) calculated that for a 90% nominal powered trial to ensure you would have at least 50% power even with a high plausible value for the variance, for a superiority trial you would need at least 9 degrees of freedom for estimating the variance.

Sim and Lewis (2011) base their recommendation about the minimum sample size for a pilot trial of 60 on ensuring a 90% chance of no more than a 10% drop in power from the nominal value (when using the UCL approach with a 90% UCL). This is different from Julious (2005) who allows a drop from an assumed power of 90% to 50% power for the high plausible value of the variance, the 95th percentile for the variance.

In the paper which aimed to demonstrate the variation in estimates taken from small samples Teare et al. (2014) suggested a total pilot trial sample size of 70. This recommendation was based on only a small gain in precision after the sample size exceeds this level and the sample size being big enough to minimise the bias in the estimation of the variability parameter.

Alternatively, we could increase the size of the main trial to accommodate the uncertainty in the variance estimate from the pilot trial, using the methods previously introduced, the UCL or NCT approach. From this perspective, using the 95% UCL method, Sim and Lewis recommend a sample size of 50 as this equates to a potential under recruitment of 30% achieving approximately 65% power in the main trial, when the nominal power is 80%, which the authors say is the lowest most investigators would be willing to accept, compared to when the pilot sample size is 30 which could lead to an under recruitment of 39% and achieving less than 60% power. Another reason Sim and Lewis recommend a sample size of 50-60 participants is based on minimising the combined sample size of the pilot and the main trial; this will be discussed in Section 3.4.4.

3.4.3 Proportional Rules of Thumb for Selecting Pilot Trial Sample Size

As previously discussed in Section 3.3, Stallard (2011) outlined a design which minimises the expected total number of patients required to lead to a successful definitive trial, a one-sided Type I error rate of 0.2 and a power requirement of 0.4. This method links back to the NETSCC definition of a pilot trial and that there must be a plan for future work; as Stallard considers the pilot trial as part of a whole programme of clinical evaluation and not as a stand-alone trial.

Stallard recommends that the sample size for a pilot trial should be approximately 3% that of the main definitive trial. As changing the value of the standardised difference will multiplicatively alter both the pilot trial sample size calculation and the calculation for the main trial by the same amount this method will always lead to the optimal pilot trial sample size being 3% of the main trial sample size. In the paper another example is given where the pilot trial sample size under this method is calculated when the standardised difference is equal to 0.2. The sample size required is 35 which is compared to the values set out by Schoenfeld (1980) and Julious (2005), 25 and 24 respectively. However, the size of the pilot recommended by the Stallard approach here is only close to these values because the standardised difference was set at 0.2, which is actually considered to be a small standardised difference according to the classifications laid out by Cohen (1992). In the paper an example is given where the standardised difference is chosen to be equal to 1. Using this to calculate the pilot trial sample size leads to a trial size of 1.4 and a main trial sample size of 42 based on using a one-sided Type I error rate of 0.025 and a power of 90% for the main trial. Additionally, for a medium standardised difference of 0.5 the Stallard approach leads to a pilot trial sample size of 6 participants and for a standardised difference of 0.8 a pilot sample size of 3; both of which are too small to be of any practical use.

Cocks and Torgerson (2013) also propose a rule based on setting the sample size of the pilot proportionate to the sample size of the main trial. They suggest that 9% of the

sample size of the main trial should be used as the sample size for the pilot trial. Their reasoning is different from that of Stallard in that it is based on using a CI approach, whereby the sample size is set such that if the observed difference in the pilot trial is zero, then the upper confidence limit should exclude the estimate that will be considered the MCID in the definitive trial. They calculate the sample size for the main trial for a range of standardised effect sizes for a main trial where the Type I error rate is 5% and the power 80%. Then the CI approach is used with a one-sided 80% confidence interval to calculate the required pilot trial sample size, by inverting the formula for the CI it follows that,

$$M = \left(\frac{Z_{1-\alpha}}{\delta - \Delta}\right)^2$$

where M is the pilot sample size, $Z_{1-\alpha}$ is the tabulated value of the standard Normal distribution of 1 minus the Type I error rate α (as discussed in Section 2.3.1), and $(\delta - \Delta)$ is the standardised effect size minus a small amount; so that the resulting CI would not include the required standardised effect size. This rule works out to be approximately 9% of the size of the required main trial sample. As opposed to Stallard this paper does recommend a minimum sample size of at least 20 participants. This minimum of 20 participants overall will be used again in Chapters 4, 5 and 7 to stop the sample size of a pilot dropping below practical levels.

The methods presented in this section are referred to as proportional external pilot trial sample sizes as they use a proportion of the main trial samples size as the sample size for the pilot trial. These rules are summarised in Table 3.3. The recommendations will be revisited in Chapter 4 to look at their effect on the power and average sample size of the trial.

Table 3.3: Proportional Rules of Thumb for External Pilot Trial Sample Size

Author	Type of Calculation	Recommended Sample Size
Stallard (2011)	Minimising the expected sample size to a successful main trial	3%
Cocks and Torgerson (2013)	CI approach to exclude MCID	9%

3.4.4 Minimising the Overall Trial Sample Size

The NETSCC define pilot trials in the context of a future trial being planned (NETSCC, 2012). Consistent with this definition is a method of choosing the sample size of the pilot trial in order to minimise the overall trial sample size i.e. the sample size of the pilot and main trial together. This is possible because the size of the pilot trial affects the size of the main trial when either the UCL approach or the NCT method is used to calculate the sample size for the main trial. If the pilot trial is large the main trial will be relatively small and if the pilot trial is small the main trial will be relatively large due to the precision around the estimate of the variance. The methods, which use this idea to choose an appropriate pilot trial sample size, are discussed here and presented in Table 3.4, the sample size presented here is the total required for a pilot trial with two treatment arms.

Table 3.4: Minimising the Overall Sample Size Rules for External Pilot Trial Sample Size (for a Two-Armed Trial)

Author	Type of Calculation	Inflation Method	Recommended Sample Size
Kieser and Wassmer (1996)	Minimising the overall trial sample size	80% UCL	20-40
Sim and Lewis (2012)	Minimising the overall trial sample size	95% UCL	≥55

In 1996 Kieser and Wassmer proposed this idea in their paper *'On the Use of the Upper Confidence Limit for the Variance from a Pilot Sample for Sample Size Determination'*. They expressed the expectation of the total sample size $N_T = M + N_M{}^*$ as,

$$\mathbb{E}(N_T) = M + \mathbb{E}(N_M{}^*), \tag{3.4}$$

where, M is the pilot trial sample size, k is the number of degrees of freedom for the variance from the pilot trial and $N_M{}^*$ is the size of the main trial dependent on the size of the pilot trial. $\mathbb{E}(N_M{}^*)$ can be estimated by $N_M k \Big/ \chi^2_{1-X,k}$ where N_M is the total main trial sample size based on a standard calculation and therefore,

$$\mathbb{E}(N_T) = M + N_M \frac{k}{\chi^2_{1-X,k}}.$$

They use this formula to calculate a pilot trial sample size to minimise the overall expected sample size by setting the size of the main trial (here, N_M) and using $X = 0.8$, where X represents the proportional upper bound of the $100X\%$ UCL for the variance as described in Chapter 2 in Browne's method. They found that a pilot sample size between 20 and 40 minimised the overall sample size of trials with a main trial sample size of between 80 and

250. That is a standardised effect size of between 0.4 and 0.7 (for 90% power based on a standard sample size calculation).

Using the same approach but with a 95% UCL for the variance estimate, Sim and Lewis (2011) found that a pilot trial of between 35 and 100 would produce the smallest combined size of the pilot trial and main RCT, for small to medium standardised effect sizes (0.2-0.6), and therefore they recommend a pilot trial sample size of greater than or equal to 55.

Both Sim and Lewis (2011) and, Kieser and Wassmer (1996) use Browne's method (1995) of inflating the variance estimate from a pilot trial. This procedure is used to allow for the imprecision in the variance estimate from a pilot trial as discussed in Chapter 2. For this procedure, Kieser and Wassmer (1996) use an 80% confidence interval as recommended by Browne (1995) whereas Sim and Lewis (2011) use a 95% confidence interval, which has the effect of increasing their estimate of the required sample size compared to Kieser and Wassmer (1996).

3.5 Summary

This chapter reviewed the literature on how to choose an appropriate sample size for an external pilot trial. Firstly, we describe why a standard powered calculation with traditional power and Type I error levels as presented in Chapter 2 may not be the correct choice; before moving on to discuss the possibility of other power and Type I error levels in the sample size calculation for an external pilot trial. The chapter then goes on to describe the differing views of many authors on how an unpowered sample size could be chosen for an external pilot trial.

The suggested levels of Type I error and power level to be used in a sample size calculation for a pilot trial are listed in Table 3.1 and the various methods set out as rules of thumb for unpowered sample size justifications are laid out in Tables 3.2, 3.3 and 3.4. The rules

presented come from different perspectives and ideas for the best way to choose a sample size for a pilot trial. The precision based methods look at the gain in precision around predicting a parameter value. The methods labelled as drop in main trial power either increase the pilot trial or the main trial sample size allowing for a fixed drop in power for the main trial if an extreme value of the variance is seen in the main trial. The approaches which result in a proportion of the main trial sample size being used as the pilot trial sample size are calculated using an approach which minimises the expected sample size of participants used in trials until a successful treatment is found (3% proportional rule) and using a CI approach to choose the sample size which would result in excluding the MCID if a difference of zero was found (9% proportional rule).

The final approach discussed, setting the pilot trial sample size in order to minimise the total sample size of the pilot and the main trial together could be argued to be the most appropriate method for publicly funded trials, as it recognises that a pilot trial is part of a larger clinical development programme; rather than a stand-alone trial. Other methods fail to recognise this issue (aside from Stallard (2011)) and thus aim to minimise both the pilot and the main trials separately which could lead to suboptimal overall combined sample sizes, this will be investigated further in Chapters 4, 5 and 7.

In Chapter 4 the work of Kieser and Wassmer (1996) and Sim and Lewis (2011), are extended by minimising the overall trial sample size that adjusts the main trial sample size based on the size of the pilot trial. Chapter 4 will describe the theoretical minimum possible overall sample size for the pilot and main trial together and therefore find the pilot trial sample size, which leads to this. The work will be extended in Chapter 5 to the issue of trial cost; by looking to minimise the financial costs of the main trial and pilot trial added together rather than the number of participants. In Chapter 7 these ideas are extended to look at how they relate to internal pilot trials that is pilot trials, where the participants are from the main trial itself.

Chapter 4

Calculations for Setting the Pilot Trial Sample Size to Minimise the Overall Sample Size

4.1 Introduction

The NETSCC state that there must be a plan for further work (among other criteria) for a trial to be labelled as a pilot (NETSCC, 2012). This future work should be considered at the planning stage of the pilot trial. There are many recommendations for setting the external pilot trial sample size as discussed in Chapter 3. This chapter will investigate which are the best methods and for different situations, assessing the methods based on the effect the pilot trial sample size has on the overall sample size. The size of the pilot affects the size of the main trial through the precision of the estimates gathered from the pilot. These are then adjusted for imprecision (for which methods were outlined in Chapter 2) and used to plan the main trial.

As discussed in Chapter 2 there have been two methods suggested to adjust the estimate of the variance for use in the planning of the main trial. The Upper Confidence Limit (UCL) method was proposed by Browne (1995) and is based on adjusting the estimate of the variance from the observed value in the pilot to its $100X\%$ UCL. The second is the Non-Central T-distribution (NCT) method (Julious and Owen, 2006), which chooses the main trial sample size to give on average the required nominal power, over all plausible values for the variance based on the pilot trial estimate and degrees of freedom.

As discussed in Chapter 3 the current methods for setting pilot trial sample size are based on a set of rules of thumb. Those referred to as flat rules of thumb are set values; fixed no matter how large the subsequent main trial will be. The rules of thumb that will be investigated further in this chapter are listed in Table 4.1. Kieser and Wassmer (1996) suggested the idea of selecting the pilot sample size in order to minimise the overall size of the pilot and main trial together. They use the UCL method with 80% UCL for the variance giving an 80% probability of achieving at least the planned power. They suggest a pilot trial sample size of 20 to 40 for main trial sample sizes between 80 and 250, which correspond to standardised effect size of 0.4 to 0.7 (for 90% power based on a standard sample size calculation). Sim and Lewis (2012) use the same method but with a 95% UCL for the variance. They calculate that a pilot trial of $n \geq 55$ would produce the smallest combined size of the pilot trial and main RCT, for small to medium standardised effect sizes (0.2 - 0.6).

Table 4.1: Flat Rules of Thumb for Pilot Trial Sample Size (for a Two-Armed Trial)

Author	Recommended Pilot Trial Sample Size
Julious (2005)	24
Kieser and Wassmer (1996)	20-40
Browne (1995)	30
Sim and Lewis (2012)	≥ 55
Teare et al. (2014)	70

As highlighted in Chapter 3 those referred to as proportional rules change dependent on the size of the subsequent main trial. The percentages listed in Table 4.2 represent the percentage of the main trial sample size to be used as the pilot trial sample size. For example, if the main trial was to have 1,500 participants the pilot trial should contain 45 people based on the 3% rule.

Table 4.2: Proportional Rules of Thumb for Pilot Trial Sample Size

Author	Recommended Pilot Trial Sample Size
Stallard (2011)	3%
Cocks and Torgerson (2013)	9%

The size of the main trial is dependent on the size of the pilot when using either of the correction methods. The effect of these methods on the expected power and sample size of the main trial and required sample size of the pilot trial is investigated in this chapter.

4.1.1 Aims

This chapter will look to extend the work of Kieser and Wassmer (1996) and Sim and Lewis (2012) by looking to:

- Minimise the overall sample size of the pilot and the main trial together using the NCT method and the UCL method,
- Find the theoretical 'optimal' values of the overall sample size, which could be achieved using these methods,
- Calculate the pilot trial sample size, which leads to this optimal value,
- Compare already existing rules of thumb (both flat and proportional) to these optimal values to assess how useful the existing methods are,
- Compare the effects of the UCL and the NCT approaches on trial sample size,
- Develop new rules of thumb based on the optimal results.

This chapter focuses on external pilot trials, which have the primary aim of estimating the variance to be used in the main trial sample size calculation. Additionally, this chapter concentrates on external pilot trials where it is assumed that there are no changes between the pilot and the following main trial, which would affect the variance estimate e.g. changing the outcome measure. In an external pilot trial, the data from the pilot sample are not included in the final analysis of the main trial. Internal pilot trials where the pilot data are included in the final analysis are considered in Chapters 6 and 7.

4.2 Minimising the Overall Sample Size Using the NCT Approach

By planning for further work after the pilot trial we can think of the pilot and main trial as one overall trial programme. It is this consideration for further work, which leads to the method of minimising the pilot, and the main trial together to produce an 'optimal' overall sample size, which is proposed in this PhD. This section uses the NCT approach as proposed by Julious and Owen (2006) to adjust the sample size for the main trial, based on the degrees of freedom for the variance estimate from the pilot trial. As such the minimum overall sample size for a variety of standardised effect sizes are generated.

4.2.1 Deriving the Minimum Overall Sample Size

To find the minimum overall sample size, the size of the pilot trial is varied over a range of values (starting at 2 per arm, to prevent the degrees of freedom from equalling zero or less, and iterating upwards) and the required main trial sample size calculated, for each standardised effect size. This adjusted main trial sample size is added to the pilot trial sample size to give the overall sample size (pilot plus main trial) n_T, through Equation 4.1,

$$n_T = m + n_M. \qquad (4.1)$$

The total sample size for a two arm trial will be denoted by N_T where $N_T = 2m + 2n_M$.

For the NCT approach the main trial sample size is calculated based on Equation 2.21. In which the required sample size n_M (the sample size for the main trial) appears on both sides of the inequality. It can be solved iteratively until the inequality is satisfied. A starting point for these iterations can be found using Equation 2.22 and rounding downwards.

Once the overall sample size for each of the pilot trial sample sizes has been calculated, the minimum overall sample size for each effect size can be found. This process is illustrated in Figure 4.1.

Figure 4.1: Process for Calculating the Minimum Overall Sample Size for the NCT Approach

Step 1: Values for Type I (α) and Type II (β) error are selected the standard deviation value (s) and the effect size (d) is set, and the starting value of m is chosen. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s was set to 1, various values of d were investigated between 0.05 and 1 and the starting value of m was chosen to be 2 participants per treatment group to prevent the degrees of freedom for the variance from being less than or equal to zero. Set $i = 1$.

Step 2: For a pilot trial sample size m_i, where i is the iteration number, estimate the main trial sample size, n_{START} from Equation 2.22.

Step 3: Using n_{START} as a starting point for n_M in Equation 2.21 iterate n_M upwards until the inequality in Equation 2.21 is satisfied.

Step 4: Estimate the overall sample size of the pilot and the main trial, n_T from Equation 4.1.

Step 5: For $i = 1$ go to Step 6, for $i > 1$ go to Step 7.

Step 6: Add 1 to the previous pilot trial sample size, m_i and i, and go to Step 2.

Step 7: If n_T for $m_i \leq n_T$ for m_{i-1} then go to Step 6. If n_T for $m_i > n_T$ for m_{i-1} then go to Step 8.

Step 8: Take n_T for m_{i-1} as n_{OPT} the minimum possible overall sample size.

For example, if the Type I error rate is chosen to be 0.05 or 5% (two-sided), the Type II error rate is set at 0.1 or 10%, if the chosen standardised effect size is 0.5 and we can start from m=2 (M=4) and set i=1. Following the algorithm to step 4 we get the following values:

Iteration	**M**	N_M	N_T
1	4	708	712

From this we can follow the rest of the algorithm. The first five loops of this would give the following results:

Iteration	**M**	N_M	N_T
1	4	708	712
2	6	334	340
3	8	264	272
4	10	236	246
5	12	220	232

The algorithm would stop when the newly calculated N_T is larger than the previous value of N_T for the chosen parameter values. In this situation the algorithm would stop at iteration 12:

Iteration	**M**	N_M	N_T
11	24	190	214
12	26	190	216

Therefore, we would say that the optimal pilot trial sample size is 24 and optimal overall trial sample size is 214 for a two armed trial.

4.2.2 Minimum Overall Sample Sizes

Table 4.3 shows the optimal overall sample size for a two-armed trial for the NCT approach for both an 80% and 90% powered main trial. The trials with higher power and a smaller standardised effect size require a larger trial sample size. Some of the sample sizes in Table 4.3 are large and dependent on the trial setting, design or funder may not be considered feasible.

Table 4.3: Minimum Overall Sample Size for the NCT Approach for Two-armed Trials

Standardised Effect Size	80% Powered Main Trial	90% Powered Main Trial
0.05	12,854	17,234
0.10	3,290	4,416
0.20	862	1,160
0.25	566	762
0.30	402	542
0.40	238	320
0.50	160	214
0.60	116	156
0.70	90	120
0.75	80	108
0.80	72	96
0.90	60	80
1.00	50	68

4.3 Minimising the Overall Sample Size Using the UCL Approach

This section uses the UCL approach to adjust the sample size for the main trial; these results will be used in Section 4.4 to compare the NCT (as seen in Section 4.2) and UCL methods.

4.3.1 Deriving the Minimum Overall Sample Size

The same approach as presented in Section 4.2 was used for finding the minimum overall sample size in this section, however instead of using the NCT approach the UCL approach was employed to adjust the main trial sample sizes.

For the UCL approach the sample size for the main trial is calculated through Equation 4.2 which is derived by replacing σ^2 in Equation 2.15 with s_{UCL}^2

$$n = \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2 \, s_{UCL}^2}{d^2}. \tag{4.2}$$

The required level of Type I error is set as well as the required standardised effect size and the probability of achieving the required power, X. Where X represents the upper confidence limit being taken for the UCL approach which gives a $100X\%$ chance of achieving the required power for the trial. This chapter investigates probability levels for X of 0.8 and 0.95. The process of calculating the minimum overall sample sizes is depicted in Figure 4.2.

Figure 4.2: Process for Calculating the Minimum Overall Trial Sample Size for the UCL Approach

Step 1: Values for Type I (α) and Type II (β) error are selected the standard deviation value (s) and the effect size (d) is set, the required level of X is chosen, and the starting value of m is chosen. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s was set to 1, various values of d were investigated between 0.05 and 1 and the starting value of m was chosen to be 2 participants per treatment group to prevent the degrees of freedom for the variance from being less than or equal to zero. Set $i = 1$.

Step 2: For a pilot trial sample size m_i, where i is the iteration number, estimate the main trial sample size, n_M from Equation 4.2.

Step 3: Estimate the overall sample size of the pilot and the main trial, n_T from Equation 4.1.

Step 4: For $i = 1$ go to Step 5, for $i > 1$ go to Step 6.

Step 5: Add 1 to the previous pilot trial sample size, m_i and i, and go to Step 2.

Step 6: If n_T for $m_i \leq n_T$ for m_{i-1} then go to Step 5. If n_T for $m_i > n_T$ for m_{i-1} then go to Step 7.

Step 7: Take n_T for m_{i-1} as n_{OPT} the minimum possible overall sample size.

For example, if the Type I error rate is chosen to be 0.05 or 5% (two-sided), the Type II error rate is set at 0.1 or 10%, if the chosen standardised effect size is 0.5 and X is chosen to be 0.8. We can start from m=2 (M=4) and set i=1. Following the algorithm to step 4 we get the following values:

Iteration	M	N_M	N_T
1	4	754	758

From this we can follow the rest of the algorithm. The first five loops of this would give the following results:

Iteration	M	N_M	N_T
1	4	754	758
2	6	408	414
3	8	330	338
4	10	294	304
5	12	274	286

The algorithm would stop when the newly calculated N_T is larger than the previous value of N_T for the chosen parameter values. In this situation the algorithm would stop at iteration 15:

Iteration	M	N_M	N_T
14	30	220	250
15	32	216	248

Therefore, we would say that the optimal pilot trial sample size is 32 and optimal overall trial sample size is 248 for a two armed trial.

4.3.2 Minimum Overall Sample Sizes

Table 4.4 shows the minimum overall sample size for a two-armed trial for the UCL approach (with both 80% and 95% confidence levels) for 80% and 90% powered main trials. Again the trials with higher powers and smaller standardised effect sizes require a higher sample size, additionally the trials, which use the 95% UCL compared to the 80% UCL need a larger sample size to maintain this higher probability level of achieving the required power.

Table 4.4: Minimum Overall Sample Size for the UCL (80 and 95%) Approaches for Two-armed Trials

	80% Powered Main Trial		90% Powered Main Trial	
Standardised Effect Size	**80% UCL**	**95% UCL**	**80% UCL**	**95% UCL**
0.05	13,762	14,444	18,266	19,092
0.10	3,632	3,912	4,796	5,134
0.20	990	1,108	1,296	1,438
0.25	658	746	858	966
0.30	474	544	616	700
0.40	284	334	368	428
0.50	194	232	248	294
0.60	142	174	182	220
0.70	110	136	140	172
0.75	100	124	126	154
0.80	90	112	112	140
0.90	74	94	94	116
1.00	64	80	80	100

4.4 Theoretical Optimal Values of Pilot Trial Sample Size

The following section uses the results found in Section 4.2 and 4.3 to find the pilot trial sample sizes, which lead to these minimum overall sample sizes. This pilot trial sample size will be referred to as the optimal pilot trial sample size, as it is the pilot trial sample size, which would lead to the theoretical minimum possible overall sample size. This process adds an additional step to the processes shown in Figures 4.1 and 4.2 as shown in Figure 4.3.

Figure 4.3: Finding the Optimal Pilot Trial Sample Size

Step 8: Take m_{i-1} as m_{opt} the pilot trial sample size that leads to the optimal overall sample size, N_{opt}.

As previously described in Section 4.2 this process was carried out for various values of the standardised effect size. Table 4.5 shows the optimal pilot trial sample size, the required main trial sample size based on this pilot trial sample size and the resulting overall sample size for a two-armed trial for all the adjustment methods. These results are calculated for 90% and 80% powered main trials with a two-sided Type I error rate of 5% and allocation ratio of 1.

Table 4.5: Theoretical Optimal Values of Pilot Trial Sample Size, Main Trial and Overall Sample Size for a Two-armed Trial for each Adjustment Method for 90% and 80% Powered Main Trials

	80% UCL Approach			95% UCL Approach			Non-Central T-distribution		
Standardised Effect Size	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**
90% Powered Main Trial									
0.05	506	17,760	18,266	794	18,298	19,092	212	17,022	17,234
0.10	210	4,586	4,796	332	4,802	5,134	108	4,308	4,416
0.20	90	1,206	1,296	144	1,294	1,438	56	1,104	1,160
0.25	70	788	858	110	856	966	44	718	762
0.30	56	560	616	90	610	700	38	504	542
0.40	40	328	368	64	364	428	30	290	320
0.50	32	216	248	50	244	294	24	190	214
0.60	26	156	182	42	178	220	20	136	156
0.70	22	118	140	36	136	172	18	102	120
0.75	20	106	126	34	120	154	16	92	108
0.80	20	92	112	32	108	140	16	80	96
0.90	18	76	94	28	88	116	14	66	80
1.00	16	64	80	26	74	100	14	54	68
80% Powered Main Trial									
0.05	420	13,342	13,762	660	13,784	14,444	148	12,706	12,854
0.10	176	3,456	3,632	278	3,634	3,912	76	3,214	3,290
0.20	76	914	990	120	988	1,108	38	824	862
0.25	58	600	658	94	652	746	32	534	566
0.30	48	426	474	76	468	544	26	376	402
0.40	34	250	284	56	278	334	20	218	238
0.50	28	166	194	44	188	232	18	142	160
0.60	22	120	142	36	138	174	14	102	116
0.70	20	90	110	30	106	136	12	78	90
0.75	18	82	100	28	96	124	12	68	80
0.80	18	72	90	28	84	112	12	60	72
0.90	16	58	74	24	70	94	10	50	60
1.00	14	50	64	22	58	80	10	40	50

It can be seen that as the standardised effect size increases the optimal pilot trial sample size decreases. The same pattern as previously seen can be observed again such that; the 95% UCL approach results in the largest pilot trial sample sizes followed by the 80% UCL and the NCT approaches respectively.

The effect of the adjustment methods can also be depicted graphically. Figures 4.4 to 4.6 show the effect on the overall sample size the adjustment methods have compared to the traditional sample size calculation, which assumes that the variance is known. The figures illustrate the results for a clinical trial with two treatment arms.

The black solid line describes a standard sample size calculation with no adjustment method applied assuming that the population variance is known, based on Equation 2.15. The red dashed curve represents the NCT method as proposed by Julious and Owen (2006). The green dotted and dashed line is the UCL method with an 80% UCL for the variance estimate. The blue dotted line is the UCL method with a 95% UCL for the variance. The overall sample sizes on the graphs are the total for a two-armed trial.

Figure 4.4: Comparing Overall Sample Sizes for each Adjustment Method and the Traditional Formula for each Pilot Trial Sample Size for a Standardised Effect Size of 0.2 for a Two-Armed Trial

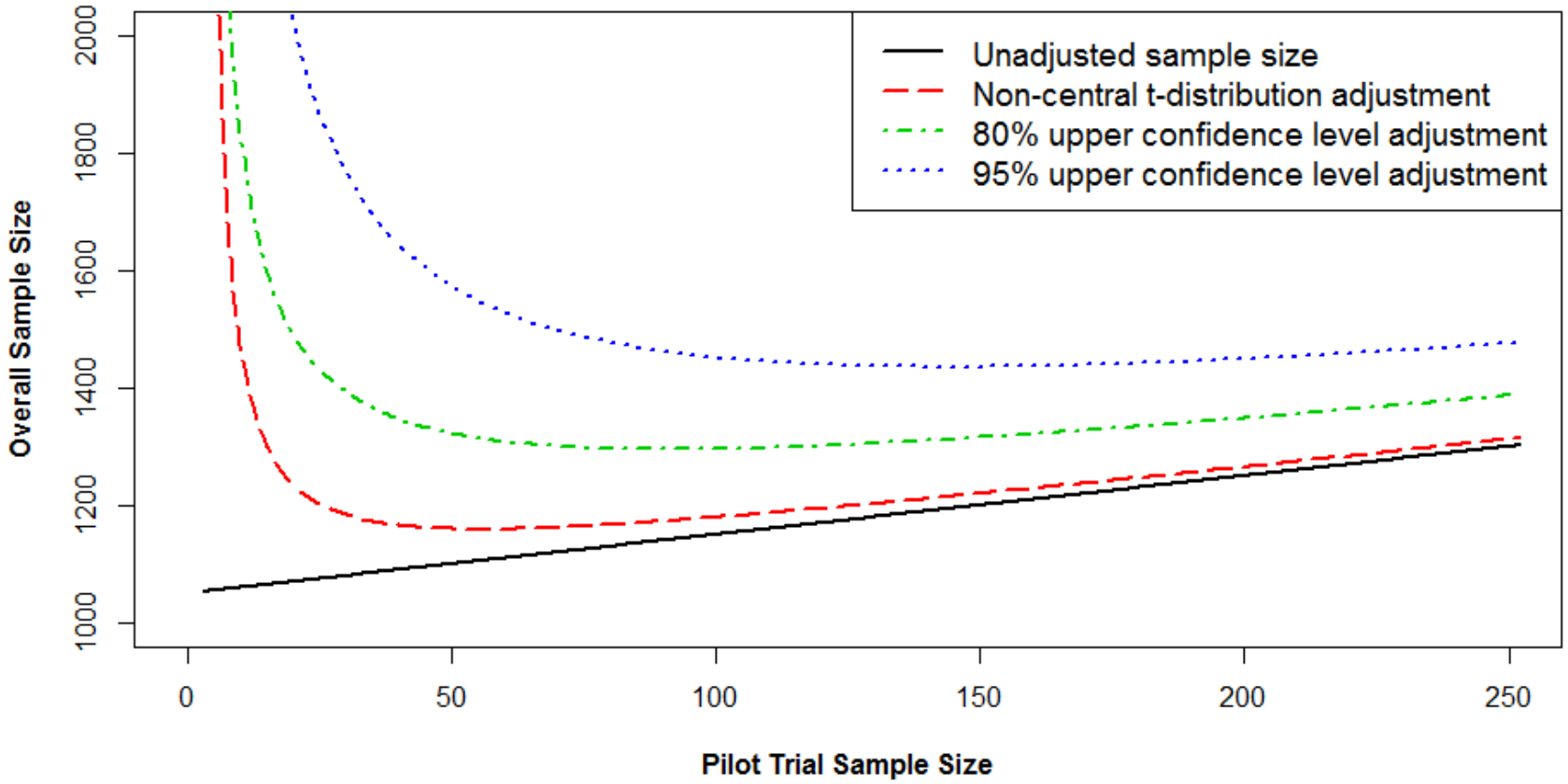


Figure 4.5: Comparing Overall Sample Sizes for each Adjustment Method and the Traditional Formula for each Pilot Trial Sample Size for a Standardised Effect Size of 0.5 for a Two-Armed Trial

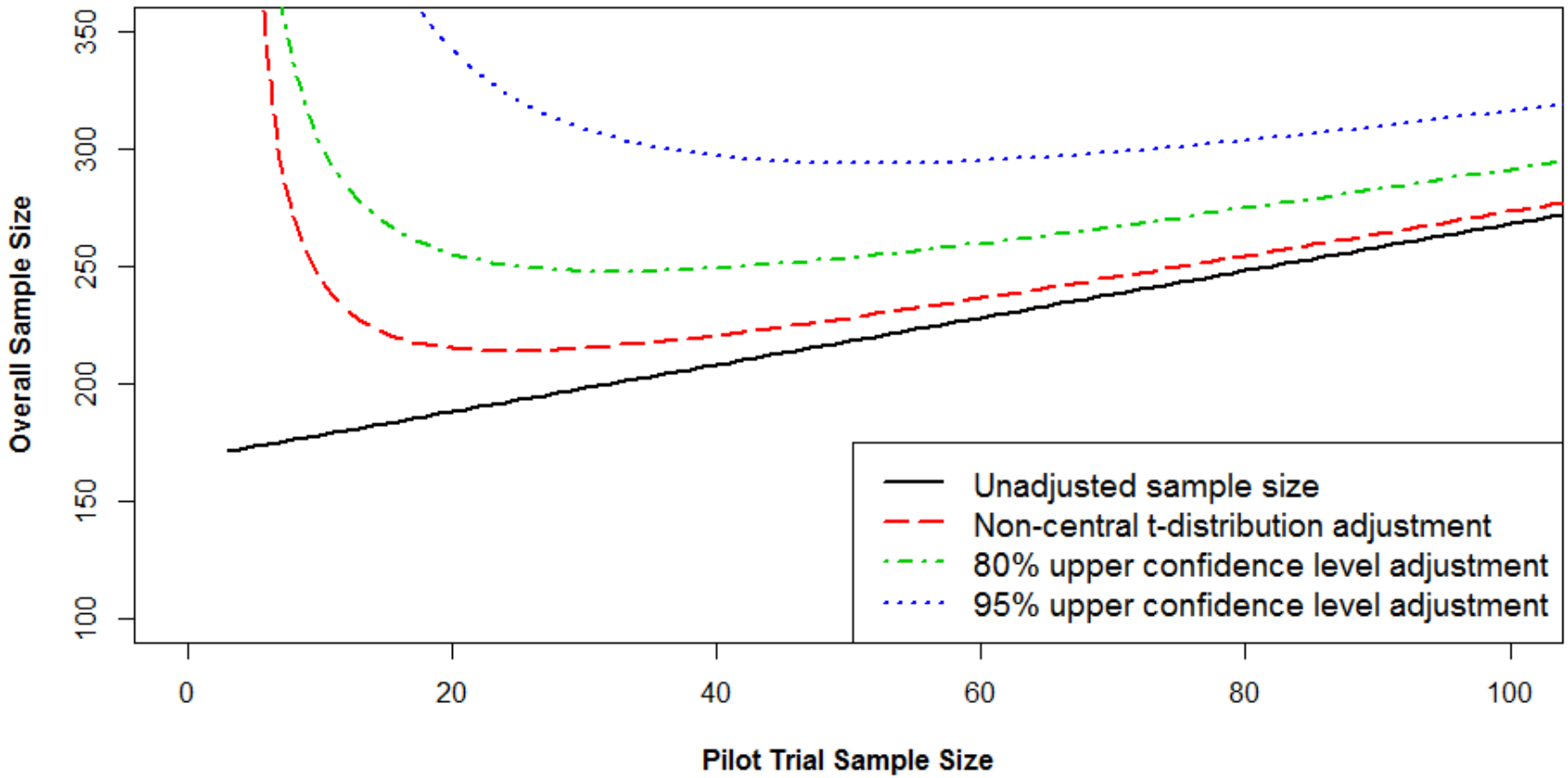
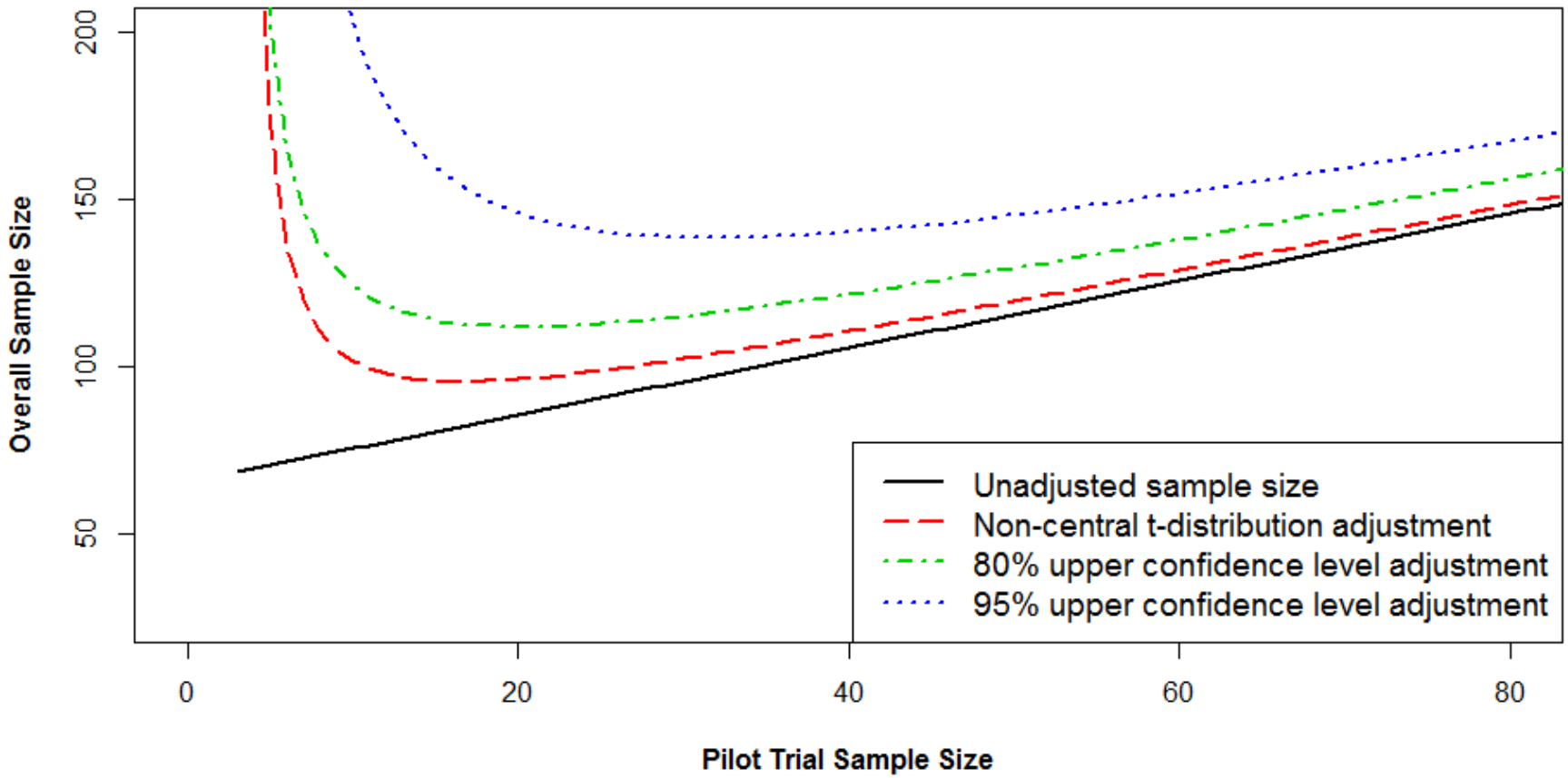


Figure 4.6: Comparing Overall Sample Sizes for each Adjustment Method and the Traditional Formula for each Pilot Trial Sample Size for a Standardised Effect Size of 0.8 for a Two-Armed Trial



It is demonstrated in Figures 4.4 to 4.6 that for the three methods, where an adjustment method has been used, as the pilot trial sample size initially increases the overall sample size decreases since the increase in pilot trial sample size is offset by a larger decrease in the main trial adjustment. However, eventually adding more participants into the pilot trial is not beneficial and the increase in pilot trial sample size is not offset by the decrease in the subsequent main trial sample size and ultimately the overall sample size begins to rise again.

There is a trade-off therefore between having a small pilot trial and a large main trial or a large pilot trial and a small main trial; the larger the pilot the more accurate the information and hence the smaller the inflation applied to the main trial sample size calculation. However, eventually the pilot will get too large and the number included in the pilot trial will outweigh the reduction in the main trial sample size. It can be seen therefore that there is a minimum possible overall sample size and, it is possible to solve the function to find the pilot trial sample size, which minimises the overall sample size, and these results are presented in Table 4.5.

Table 4.5 shows the results from the graphs in Figures 4.4 to 4.6 for 90% powered main trials, plus the results for 80% powered main trials, displaying the pilot sample size for which the overall sample size is minimised, the resulting main trial sample size based on this pilot trial sample size and the minimum possible overall sample size. These results will be referred to as the optimal values as these are the lowest numbers you could theoretically achieve and still on average have the required power for the trial.

4.5 Comparing the Optimal Values to the Flat Rules of Thumb

In the previous section Figures 4.4 to 4.6 were used to compare the resulting overall sample size of three types of variance adjustment to the unadjusted overall sample size, which assumes that the variance is known. In this section the flat rules of thumb

presented in Chapter 2 are added to the plots and compared to the theoretical 'optimal' value of pilot trial sample size (Figures 4.7 to 4.9).

Figure 4.7: Comparing Overall Sample Sizes for each Correction Method for Varying Pilot Trial Sample Sizes for a Standardised Difference of 0.2 for a Two-Armed Trial

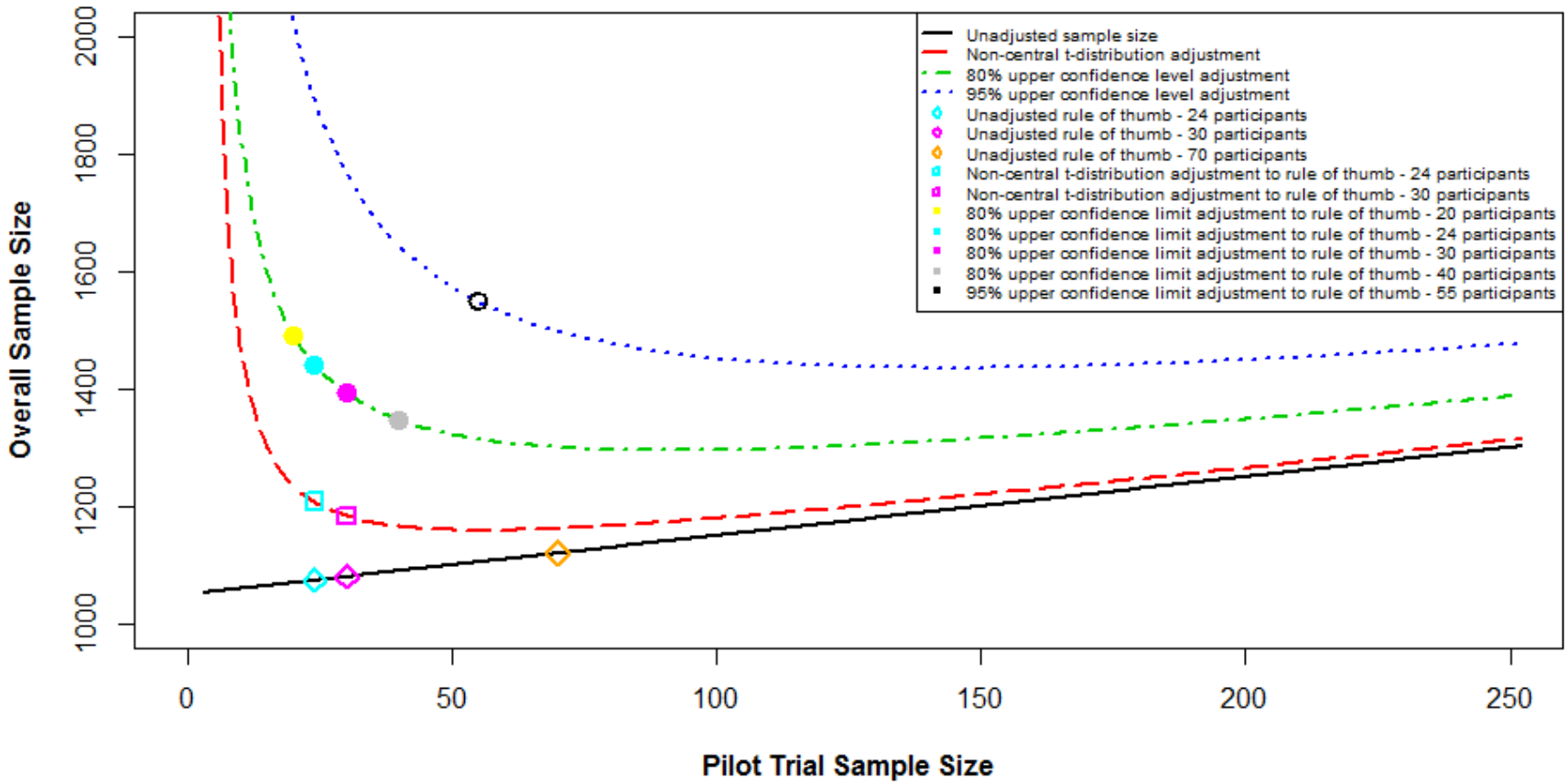


Figure 4.8: **Comparing Overall Sample Sizes for each Correction Method for Varying Pilot Trial Sample Sizes for a Standardised Difference of 0.5 for a Two-Armed Trial (Sample Size Total for a Two-armed Trial)**

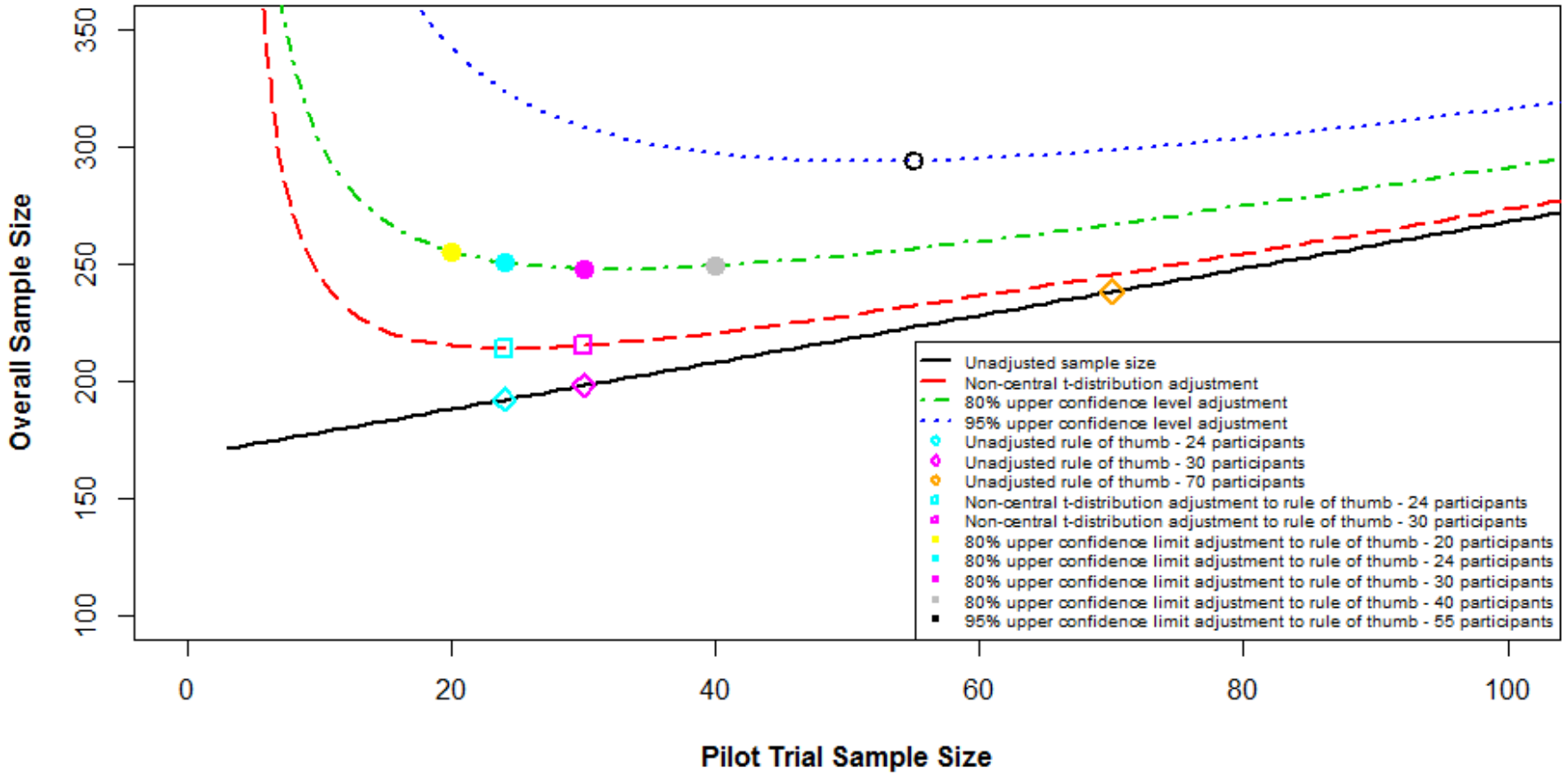
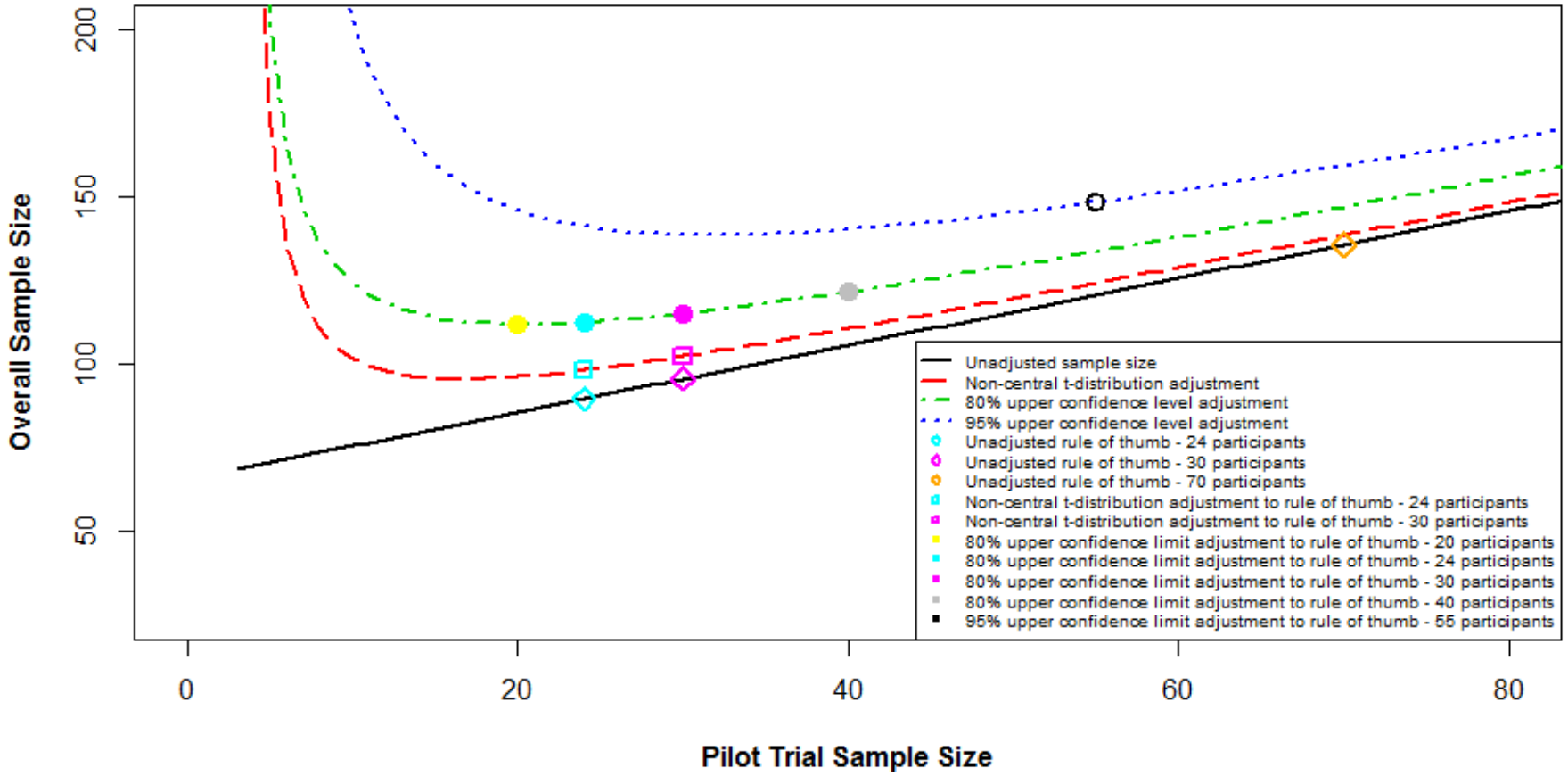


Figure 4.9: Comparing Overall Sample Sizes for each Correction Method for Varying Pilot Trial Sample Sizes for a Standardised Difference of 0.8 for a Two-Armed Trial (Sample Size Total for a Two-armed Trial)



The black solid line again depicts a standard sample size calculation with no adjustment applied hence assuming that the population variance is known. The points on the line show the resulting overall sample size if the rules of thumb 24, 30 or 70 were used with no adjustment method applied. The red dashed curve represents the NCT method. The points show the resulting overall sample size of the rules of thumb if 24 or 30 subjects were used for the pilot trial. The green dotted and dashed line is the UCL method with an 80% UCL for the variance. The points represent the rules of thumb of 20 and 40 as set out by Kieser and Wassmer (1996) as well as 24 or 30 rules. The blue dotted line is the UCL method with a 95% UCL for the variance. The point for a pilot trial sample size of 55 has been added here, as this was the sample size recommended by Sim and Lewis (2012) to minimise the overall sample size. The overall sample sizes on the graphs are the total for a two-armed trial.

The following table (Table 4.6) displays differences in the number of participants between the theoretical optimal values and the inflation methods using the rules of thumb for both the pilot trial sample size and the overall sample size used for a 90% powered main trial. A negative value indicates the rule of thumb uses fewer participants than the optimal value indicates. The results for 80% powered main trials are presented in Table 4.7.

Table 4.6: Distances from Optimal Values for the Rules of Thumb for Varying Standardised Differences for a Main Trial Power of 90% Based on a Two Armed Trial

Standardised Effect Size	Pilot Trial Sample Size	Overall Trial Sample Size	Correction Method	Distance from Optimal Pilot Trial Sample Size	Distance from Optimal Overall Trial Sample Size
0.2	**20**	**1,492**	**80% UCL**	-70	196
	24	**1,442**	**80% UCL**	-66	146
	30	**1,394**	**80% UCL**	-60	98
	40	**1,350**	**80% UCL**	-50	54
	55	**1,549**	**95% UCL**	-89	111
	24	**1,208**	**NCT**	-32	48
	30	**1,186**	**NCT**	-26	26
0.5	**20**	**256**	**80% UCL**	-12	8
	24	**252**	**80% UCL**	-8	4
	30	**250**	**80% UCL**	-2	2
	40	**250**	**80% UCL**	+8	2
	55	**295**	**95% UCL**	+5	1
	24	**214**	**NCT**	0	0
	30	**216**	**NCT**	+6	2
0.6	**20**	**184**	**80% UCL**	-6	2
	24	**182**	**80% UCL**	-2	0
	30	**182**	**80% UCL**	+4	0
	40	**186**	**80% UCL**	+14	4
	55	**221**	**95% UCL**	+13	1
	24	**158**	**NCT**	+4	2
	30	**160**	**NCT**	+10	4
0.8	**20**	**112**	**80% UCL**	0	0
	24	**114**	**80% UCL**	+4	2
	30	**116**	**80% UCL**	+10	4
	40	**122**	**80% UCL**	+20	10
	55	**149**	**95% UCL**	+23	9
	24	**100**	**NCT**	+8	4
	30	**104**	**NCT**	+14	8

From Table 4.6 and Figures 4.7 to 4.9 we can see that for medium standardised effect sizes (0.5 and 0.6) the suggested rules of thumb are very close to the optimal pilot sample size. However, when the standardised effect size moves away from these values the rules of thumb are less useful. For small standardised effect sizes (e.g. 0.2) the rules of thumb underestimate the required size of the pilot by as many as 89 participants. For large standardised effect sizes (e.g. 0.8) the rules of thumb overestimate the number of participants required by as much as 23. These results indicate that the larger the main trial the larger the pilot trial should be in order to minimise the overall sample size; therefore, one fixed flat pilot trial sample size will not be suitable for all trials.

In relation to the overall sample size, overestimating the pilot sample size is not as costly as underestimating in terms of over recruitment of participants, as demonstrated in the graphs given that the slope of the right hand side of the graph is flatter than the left hand side. Therefore, for the same change in pilot trial sample size overestimation compared to underestimation the change in overall sample size will be comparatively less. Consequently, for larger standardised effect sizes the difference of the rules of thumbs to the optimal size is lower than for the smaller standardised effect sizes.

It can be seen that as the standardised effect size increases the effect of using a suboptimal pilot trial sample size decreases. This is due to the smaller numbers already involved in the trial due to the larger standardised effect size.

For a small standardised effect size of 0.2 using a pilot of 55 participants and the 95% UCL method could lead to an over recruitment beyond the theoretical minimum for this approach of 111 additional participants to the overall sample size. Using a pilot trial sample size of 20 and the 80% UCL method could lead to increasing the overall sample size by up to 196 participants over the theoretical optimal overall sample size for this method. If the NCT method is used after a pilot trial with a sample size of 24 this could result in an extra 48 people than needed being recruited in the optimal design. For large effect sizes these are 9, 10 and 8 extra participants for the 95% UCL method (with a pilot

trial sample size of 55), the 80% UCL method (with a pilot trial sample size of 40) and the NCT method (with a pilot trial sample size of 30) respectively. It can also be seen that the NCT approach produces consistently lower overall sample sizes than any of the other methods.

Table 4.7: Distances from Optimal Values for the Rules of Thumb for Two Armed Trials for Varying Standardised Differences for a Main Trial Power of 80% Based on a Two Armed Trial

Standardised Effect Size	Pilot Trial Sample Size	Overall Trial Sample Size	Correction Method	Distance from Optimal Pilot Trial Sample Size	Distance from Optimal Overall Trial Sample Size
0.2	**20**	**1,120**	**80% UCL**	-56	130
	24	**1,084**	**80% UCL**	-52	94
	30	**1,050**	**80% UCL**	-46	60
	40	**1,018**	**80% UCL**	-36	28
	55	**1,171**	**95% UCL**	-65	63
	24	**874**	**NCT**	-14	12
	30	**866**	**NCT**	-8	4
0.5	**20**	**196**	**80% UCL**	-8	2
	24	**194**	**80% UCL**	-4	0
	30	**194**	**80% UCL**	2	0
	40	**198**	**80% UCL**	12	4
	55	**235**	**95% UCL**	11	3
	24	**162**	**NCT**	6	2
	30	**166**	**NCT**	12	6
0.6	**20**	**144**	**80% UCL**	-2	2
	24	**142**	**80% UCL**	2	0
	30	**144**	**80% UCL**	8	2
	40	**150**	**80% UCL**	18	8
	55	**179**	**95% UCL**	19	5
	24	**120**	**NCT**	10	4
	30	**124**	**NCT**	16	8
0.8	**20**	**90**	**80% UCL**	2	0
	24	**92**	**80% UCL**	6	2
	30	**94**	**80% UCL**	12	4
	40	**102**	**80% UCL**	22	12
	55	**125**	**95% UCL**	27	13
	24	**78**	**NCT**	12	6
	30	**84**	**NCT**	18	12

Table 4.7 reflects the same pattern for the 80% powered main trial as was seen for the 90% powered main trial. It can be seen that for medium standardised effects sizes again the suggested rules of thumb are very close to the optimal pilot trial sample sizes.

For small standardised effect sizes, the rules of thumb still underestimate the required size of the pilot as with the 90% powered trials however, because in an 80% powered trial the main trial will be smaller the underestimation is lower due to the increase in the size of the pilot relative to the size of the main trial. For large standardised effect sizes, the rules of thumb overestimate how large the pilot trial needs to be, for 80% powered trials this is worse because as previously mentioned the relative size of the pilot to the main trial is larger than for a 90% powered trial.

It may be noted that for large values of the standardised effect size the suggested pilot trial sample size falls to a level which may be considered too low to achieve the other objectives of a pilot trial (as outlined in Chapter 1). This is because pilot trials are not only used to estimate the variance of the outcome measure, but also to assess objectives such as testing the feasibility of trial processes or predicting the likely dropout rate. We must consider these other objectives as well as more practical considerations. For these reasons the suggestion would be not to use a pilot trial sample size below 10 per arm or 20 for a two-armed trial, as this is the lowest of all the flat rules of thumb presented in Table 4.1. The following table (Table 4.8) represents the optimal results with a cap on the lower limit of the pilot trial sample size of 10 per treatment group or 20 overall for a two group trial.

Table 4.8: Theoretical Optimal Values of Pilot Trial Sample Size, Main Trial and Overall Sample Size for a Two-armed Trial for each Adjustment Method for 90% and 80% Powered Main Trials with a Cap on the Lower Limit of Pilot Trial Sample Size at 10 participants

	80% UCL			95% UCL			Non-Central T-distribution		
Standardised Effect Size	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**
90% Powered Main Trial									
0.05	506	17,760	18,266	794	18,298	19,092	212	17,022	17,234
0.10	210	4,586	4,796	332	4,802	5,134	108	4,308	4,416
0.20	90	1,206	1,296	144	1,294	1,438	56	1,104	1,160
0.25	70	788	858	110	856	966	44	718	762
0.30	56	560	616	90	610	700	38	504	542
0.40	40	326	368	64	364	428	30	290	320
0.50	32	216	248	50	244	294	24	190	214
0.60	26	156	182	42	178	220	20	136	156
0.70	22	118	140	36	136	172	20	100	120
0.75	20	106	126	34	120	154	20	88	108
0.80	20	92	112	32	108	140	20	78	98
0.90	20	74	94	28	88	116	20	62	82
1.00	20	60	80	26	74	100	20	50	70
80% Powered Main Trial									
0.05	420	13,342	13,762	660	13,784	14,444	148	12,706	12,854
0.10	176	3,456	3,632	278	3,634	3,912	76	3,214	3,290
0.20	76	914	990	120	988	1,108	38	824	862
0.25	58	600	658	94	652	746	32	534	566
0.30	48	426	474	76	468	544	26	376	402
0.40	34	250	284	56	278	334	20	218	238
0.50	28	166	194	44	188	232	20	140	160
0.60	22	120	142	36	138	174	20	98	118
0.70	20	90	110	30	106	136	20	72	92
0.75	20	80	100	28	96	124	20	62	82
0.80	20	70	90	28	84	112	20	56	76
0.90	20	56	76	24	70	94	20	44	64
1.00	20	44	64	22	58	80	20	36	56

In Table 4.8 a lower cap is placed on the pilot trial sample size so that it cannot drop lower than the recommended minimum of 10 participants per treatment arm or 20 participants in total. From Table 4.8 (90% powered main trials) it can be seen that for the 80% UCL method the optimal pilot trial sample size becomes 10 when the standardised difference reaches 0.8 or higher. For the 95% UCL method the optimal pilot trial sample size never falls as low as 10 so the cap does not come into effect here. For the NCT method the optimal pilot trial sample size reverts to being 10 once the standardised difference reaches 0.7 or higher. For 80% powered main trials the caps can be seen to come into effect earlier for the 80% UCL method when the standardised difference reaches 0.7 and for the NCT method when it reaches 0.5 or more. The cap still has no effect on the results from the 95% UCL method.

It should be noted that although the exact calculation for the NCT approach Equation 2.21 has been used here to gain the most accurate results, in practice using the approximation in Equation 2.22 will result in an overall sample size of one subject less than the exact calculation at the most as seen in Julious and Owen (2006)

4.6 Comparing the Optimal Values to the Proportional Rules of Thumb

This section will extend the work of Stallard (2011), whose paper aims to minimise the expected total number of patients required to lead to a successful definitive trial and recommends a pilot trial sample size of 3% of the main trial sample size. The objective in this section is to compare this proportional rule of thumb to the flat rules of thumb presented earlier.

In this section an algorithm is derived which finds the main trial sample size based on a chosen proportion for pilot trial sample size and standardised effect size; given the restriction that the pilot trial should be a fixed proportion of the main trial sample size. From this a pilot trial and main trial sample size is derived for varied standardised effect

sizes. Other proportions than 3% are also investigated. The proportions are investigated to find out which work well in which situations and if there is a proportion which works in all situations.

4.6.1 Deriving the Optimal Pilot Trial Sample Size Methods

In order to calculate the required number of participants for the pilot trial knowledge of the required number for the main trial is needed. The argument becomes circular because in order to calculate the required number for the main trial you need to know the required number for the pilot based on using one of the inflation methods for the variance.

To solve the problem, the following procedure was designed and implemented. The initial main trial sample size (n_M) was calculated based on the unadjusted formula (Equation 2.15) the required pilot trial sample size was calculated from this by multiplying n_M by p, the required sample size proportion for the pilot trial sample. This pilot trial sample size (m) was then used to calculate the required n_M and so on, as described in the following algorithm (Figure 4.10) and displayed in Figure 4 .11.

Figure 4.10: Algorithm to Calculate Pilot Trial Sample Size to Minimise Overall Trial Sample Size Based on Proportional Methods of Setting the Pilot Trial Sample Size

Step 1: Calculate the main trial sample size n_1 from Equation 2.15 based on the chosen standardised effect size

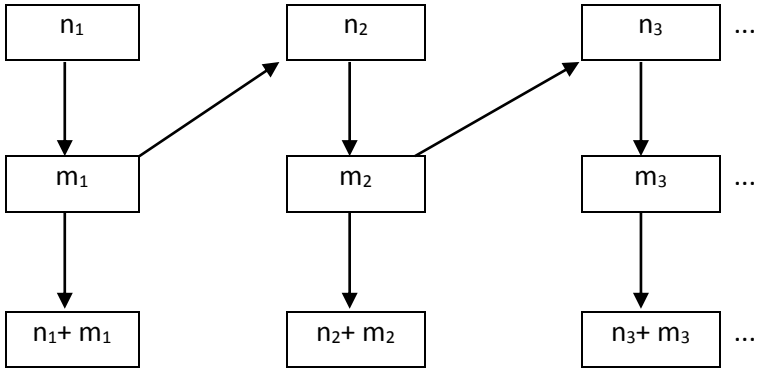
Step 2: Calculate m_1 by multiplying n_1 by p

Step 3: Calculate n_2 by using m_1 in either the UCL method or the NCT method

Step 4: Calculate m_2 by multiplying n_2 by p

Step 5: Repeat steps 3 and 4 until the values converge

Figure 4 .11: Flow Diagram Showing the Algorithm for the Proportional Approach



During the implementation of this approach a problem was identified. If the required pilot trial sample size based on the rule fell below 2 the algorithm could no longer continue as the degrees of freedom falls to zero or less. In order to combat this issue, it was decided that a lower limit for the pilot trial sample size should be set. A pilot trial sample size of 20 participants was chosen as the limit as in the flat rules of thumb section, due to this being the lowest recommended sample size in Table 4.1.

4.6.2 The 3% Rule

The following table, Table 4.9, shows the results from the procedure for the pilot trial and the main trial sample size based on the pilot trial being 3% of the size of the main trial, for a given standardised difference.

The restrictive nature of setting the pilot to 3% of the size of the main trial has resulted in an overall sample size difference of 56 (1,352 from Table 4.9 compared to 1,296 from

Table 4.4) between this method and method which allows the pilot trial sample size to take any value, for the 80% UCL method, a main trial power of 90% and with a standardised effect size of 0.2. The increase in the overall sample sizes for the standardised differences of 0.5, 0.6 and 0.8 are 4, 1 and 0 respectively. These numbers are not as large as they could have been because of the cap on the lower limit for pilot trial sample size having been set at 20 participants for the 3% method which is close to the optimal pilot trial sample size for the 80% UCL method for standardised differences of 0.6 and 0.8.

For the NCT method the 3% method has also resulted in a marginal increase in the required overall number of participants. For the standardised difference of 0.2 the increase is 20, again the lower limit of 20 for the pilot trial from the 3% method means that there is little or no difference for the standardised differences of 0.5, 0.6 and 0.8; 2, 0, 0 respectively. A similar pattern can be seen for a main trial power of 80%. These results again emphasise that rules of thumb are suitable when the standardised difference is >0.5 but for small standardised differences may result in an over recruitment.

Table 4.9: Pilot Trial, Main Trial and Overall Sample Size for a Two-Armed Trial Based on the 3% Rule with 90% or 80% Power and 5% Type I Error Rate in Main Trial

	80% UCL			95% UCL			Non-Central T-distribution		
Standardised Effect Size	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**
90% Powered Main Trial									
0.05	534	17,736	18,270	560	18,610	19,170	508	16,900	17,408
0.10	142	4,682	4,824	154	5,132	7,286	130	4,294	4,424
0.20	40	1,312	1,352	48	1,550	1,598	36	1,150	1,186
0.25	28	890	918	34	1,086	1,120	24	778	802
0.30	20	654	674	26	824	850	20	542	562
0.40	20	368	388	20	504	524	20	306	326
0.50	20	236	256	20	324	344	20	196	216
0.60	20	164	184	20	224	244	20	136	156
0.70	20	122	142	20	166	186	20	100	120
0.75	20	106	126	20	144	164	20	88	108
0.80	20	92	112	20	126	146	20	78	98
0.90	20	74	94	20	100	120	20	62	82
1.00	20	60	80	20	82	102	20	50	70
80% Powered Main Trial									
0.05	402	13,362	13,764	424	14,120	14,544	380	12,618	12,998
0.10	108	3,560	3,668	120	3,952	4,072	96	3,200	3,296
0.20	32	1,016	1,048	38	1,226	1,264	26	852	878
0.25	22	698	720	28	870	898	20	554	574
0.30	20	490	510	22	668	690	20	386	406
0.40	20	276	296	20	378	398	20	218	238
0.50	20	176	196	20	242	262	20	140	160
0.60	20	124	144	20	168	188	20	98	118
0.70	20	90	110	20	124	144	20	72	92
0.75	20	80	100	20	108	128	20	62	82
0.80	20	70	90	20	96	116	20	56	76
0.90	20	56	76	20	76	96	20	44	64
1.00	20	44	64	20	62	82	20	36	56

4.6.3 Other Proportional Pilot Trial Rules

As previously discussed pilot trials should increase in size as the sample size of the main trial increases. This is one of the flaws of flat rules of thumb. Therefore, the idea of proportional pilot trials is worthy of further investigation. A recent paper by Cocks and Torgerson (2013) suggested a proportional pilot of 9% of the main trial. This section will investigate this and other levels of pilot to main trial proportionality in the same way as described for the 3% level in the previous section.

The following table (Table 4.10) shows the results of pilot trial, main trial and minimum overall sample size for varying proportionality levels for the pilot trial versus the main trial. The algorithm, shown in Figure 4.10 was used again and applied in the same way as in Section 4.6.2.

Table 4.10: Pilot Trial Sample Size and Overall Sample Size for a Two-Armed Trial Based on Varying Proportions of the Main Trial as the Pilot Trial Sample Size for the 80% UCL Correction and the NCT Method for a 90% Powered Main Trial

	80% UCL			95% UCL			Non-Central T-distribution		
Standardised Effect Size	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**	**Pilot**	**Main**	**Overall**
5% proportional pilot trial									
0.2	64	1,246	1,310	72	1,424	1,496	56	1,108	1,164
0.5	20	236	256	20	324	344	20	196	216
0.8	20	92	112	20	126	146	20	78	98
9% proportional pilot trial									
0.2	108	1,192	1,300	120	1,322	1,442	98	1,082	1,180
0.5	22	234	256	28	292	320	20	196	216
0.8	20	92	112	20	126	146	20	78	98
10% proportional pilot trial									
0.2	120	1,184	1,304	132	1,308	1,440	108	1,080	1,188
0.5	24	230	254	30	284	314	20	196	216
0.8	20	92	112	20	126	146	20	78	98
20% proportional pilot trial									
0.2	230	1,142	1,372	246	1,228	1,474	214	1,066	1,280
0.5	42	210	252	50	246	296	38	184	222
0.8	20	92	112	24	118	142	20	78	98
50% proportional pilot trial									
0.2	554	1,108	1,662	582	1,162	1,744	530	1,058	1,588
0.5	96	192	288	108	216	324	88	176	264
0.8	42	82	124	48	96	144	36	72	108

Although the proportional methods are an improvement on the flat rules of thumb in that they allow the pilot to be larger for large main trials and smaller for small main trials, they still place a restriction on the sample size of the pilot trial, which means that the optimal overall sample size may not be achieved.

To look at which proportions would be optimal for this kind of rule of thumb Table 4.5 was used to calculate what percentage of the main trial the optimal pilot trial would be. Using the optimal results presented in Table 4.5, the optimal proportional pilot trial sample size was calculated by dividing the optimal pilot trial sample size by the main trial sample size to give a proportion. These results can be seen in Table 4.11.

Table 4.11: Optimal Proportional Pilot Trial Sample Sizes for a Two-Armed Trial for Main Trial Sample Sizes with 80% and 90% Power

	80% UCL			95% UCL			Non-Central T-distribution		
Standardised Effect Size	**Pilot**	**Main**	**Proportion**	**Pilot**	**Main**	**Proportion**	**Pilot**	**Main**	**Proportion**
90% Powered Main Trial									
0.05	506	17,760	0.03	794	18,298	0.04	212	17,022	0.01
0.10	210	4,586	0.05	332	4,802	0.07	108	4,308	0.03
0.20	90	1,206	0.07	144	1,294	0.11	56	1,104	0.05
0.25	70	788	0.09	110	856	0.13	44	718	0.06
0.30	56	560	0.10	90	610	0.15	38	504	0.08
0.40	40	328	0.12	64	364	0.18	30	290	0.10
0.50	32	216	0.15	50	244	0.20	24	190	0.13
0.60	26	156	0.17	42	178	0.24	20	136	0.15
0.70	22	118	0.19	36	136	0.26	20	100	0.20
0.75	20	106	0.19	34	120	0.28	20	88	0.23
0.80	20	92	0.22	32	108	0.30	20	78	0.26
0.90	20	74	0.27	28	88	0.32	20	62	0.32
1.00	20	60	0.33	26	74	0.35	20	50	0.40
80% Powered Main Trial									
0.05	420	13,342	0.03	660	13,784	0.05	148	12,706	0.01
0.10	176	3,456	0.05	278	3,634	0.08	76	3,214	0.02
0.20	76	9,14	0.08	120	988	0.12	38	824	0.05
0.25	58	600	0.10	94	652	0.14	32	534	0.06
0.30	48	426	0.11	76	468	0.16	26	376	0.07
0.40	34	250	0.14	56	278	0.20	20	218	0.09
0.50	28	166	0.17	44	188	0.23	20	140	0.14
0.60	22	120	0.18	36	138	0.26	20	98	0.20
0.70	20	90	0.22	30	106	0.28	20	72	0.28
0.75	20	80	0.25	28	96	0.29	20	62	0.32
0.80	20	70	0.29	28	84	0.33	20	56	0.36
0.90	20	56	0.36	24	70	0.34	20	44	0.45
1.00	20	44	0.45	22	58	0.38	20	36	0.56

It can be seen that no proportion is optimal for all standardised effect sizes in terms of minimising the overall sample size. As with the results looking at the actual numbers involved the NCT method results in smaller proportions and hence smaller pilots than the other approaches. The resulting proportions increase as the standardised effect size increases. The main trial sample size is proportional to $1/d^2$. Therefore when the pilot trial sample size is divided by the main trial sample size to calculate the proportion, the pilot sample size is multiplied by d^2 consequently as the effect size increases the resulting proportions increase. For the NCT approach, the optimal proportion for a standardised effect size of 0.05 is 1% whereas for a standardised effect size of 1 the optimal proportion is 40%, for a 90% powered main trial.

4.7 The Effect of Using the NCT Approach

The 80% UCL method requires fewer trial participants than the 95% UCL method since the 80% UCL method gives less of a chance of achieving the required power. The NCT method consistently requires even fewer subjects for the same fixed parameters. This section explores and compares the methods; how they affect the sample size of a trial and their effect on the power of the trial, in order to investigate the cost in terms of power of using the NCT approach over the UCL approaches.

4.7.1 Inflation Factors

In the paper by Julious and Owen (2006) inflation factors are presented which, if multiplied by the sample size given by the standard calculation (Equation 2.15), will result in the same sample size as if the calculation was conducted using the NCT approach (Equation 2.22). The inflation factor can be calculated from

$$IF = \frac{\left(2\left[tinv\left(1-\beta,\ k,\ Z_{1-\alpha/2}\right)\right]^2 s^2\right)/d^2}{\left(2\left(Z_{1-\beta}+\ Z_{1-\alpha/2}\right)^2 s^2\right)/d^2}$$

$$IF = \frac{\left[tinv\left(1-\beta,\ k,\ Z_{1-\alpha/2}\right)\right]^2}{\left(Z_{1-\beta}+\ Z_{1-\alpha/2}\right)^2}.$$

These inflation factors depend on the pilot trial sample size, the Type I error rate and the Type II error rate. Table 4.12 shows inflation factors based on the NCT approach for varying pilot trial sample sizes for a two-sided Type I error rate of 5% and power requirements of 80 and 90%.

Table 4.12: Inflation Factors for the Sample Size Calculation for the NCT Approach when the Type I Error is 5%

	Power	
Pilot Trial Sample Size	**90%**	**80%**
20	1.156	1.099
24	1.125	1.080
30	1.097	1.062
40	1.071	1.045
50	1.055	1.036
70	1.039	1.025
100	1.027	1.017
200	1.013	1.009

These inflation factors can also be calculated for the UCL approach. The inflation factor represents how much larger the adjusted sample size is compared to the standard calculation therefore; they are calculated by dividing Equation 4.2 by Equation 2.15, with σ^2 replaced by s^2. Equation 4.2 is Equation 2.15 where σ^2 is replaced by s^2_{UCL}, where the value of s^2_{UCL} can be found using Equation 2.12. This can be seen in Equation 4.3 where k is the degrees of freedom for the variance estimate from the pilot trial and IF, is the inflation factor.

$$IF = \frac{\left(2\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2\left(k/\chi^2_{1-X,k}\right)s^2\right)/d^2}{\left(2\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 s^2\right)/d^2}$$

$$IF = \frac{k}{\chi^2_{1-X,k}} \tag{4.3}$$

The inflation factors for the UCL approach can be seen to depend only on the sample size of the pilot trial through the degrees of freedom for the variance estimate and the chosen level of X, the chosen level of confidence. Table 4.13 shows inflation factors for the UCL approach for varying pilot trial sample size and chosen level of X, either 0.8 or 0.95, relating to 80% and 95% UCL approaches respectively. No power or Type I error rate needs to be selected to calculate the inflation factor for the UCL approach, as can be seen above in Equation 4.3, the inflation factor for the UCL approach does not depend on the power or the Type I error rate.

Table 4.13: Inflation Factors for the Sample Size Calculation Using the UCL Approach

Pilot Trial Sample Size	80% UCL	95% UCL
20	1.400	1.917
24	1.349	1.783
30	1.297	1.654
40	1.244	1.527
50	1.211	1.450
70	1.172	1.359
100	1.139	1.287
200	1.093	1.190

Table 4.14 investigates for which level of X would the UCL method approaches the NCT method in terms of the same main trial sample size; the resulting inflation factors are also reported. It can be seen that as the pilot trial sample size increases the value of X which equates the methods tends to 0.5 and the inflation factor tends to 1.

Table 4.14: Inflation Factors and Levels of X for the UCL Approach that give the same Sample Size as the NCT Approach

Pilot Trial Sample Size	Confidence Level Proportion (X)	Inflation Factor
90% Powered Main Trial		
20	0.622	1.156
24	0.611	1.125
30	0.599	1.097
40	0.586	1.071
50	0.577	1.056
70	0.565	1.039
100	0.554	1.027
200	0.538	1.013
80% Powered Main Trial		
20	0.566	1.099
24	0.560	1.080
30	0.553	1.062
40	0.546	1.045
50	0.541	1.036
70	0.534	1.025
100	0.529	1.017
200	0.520	1.008

4.7.2 Power Simulations

It can be seen from Tables 4.12 and 4.13 that the inflation factors for the NCT approach are consistently smaller than those for the UCL approach for the same power and alpha levels. Hence as observed in the previous section, the NCT approach leads to smaller main trial sample sizes for the same pilot trial sample size compared to the UCL approach. This section looks at the cost in terms of power of using the NCT approach compared to the UCL approaches.

To investigate the effect of using the NCT approach as opposed to the UCL methods simulations were carried out to compute the average power of trials, when using the adjustment methods. A pilot trial was simulated with two treatment arms – one control arm and one experimental treatment arm. For the control arm the results were drawn from a Normal distribution with a mean of 0 and a variance of 1. The experimental arm results were drawn from a Normal distribution with a mean equal to the required effect size and a variance of 1. Depending on the adjustment method under investigation the pilot trial sample size was set to the optimal value for that approach and chosen effect size. The pilot trial data was then used to produce an estimate of the variance; this estimate was then used to calculate the sample size required for the main trial (according to the adjustment method selected). The main trial sample size was based on a Type I error rate of 5%, a Type II error rate of 10% and equal allocation between the treatment groups.

Using the same method as previously detailed for the pilot trial the results for the main trial were generated using the sample size estimated. This simulation was repeated 10,000 times for each situation. The results of the main trial were analysed using a t-test and hence the power of the design was estimated by calculating the number of the simulated trials, which rejected the null hypothesis of no difference between the groups. The results of these simulations can be seen in Table 4.15.

Figure 4.12: Process for the Simulation Study Looking at Average Power

Step 1: Values for Type I (α) and Type II (β) error are selected the standard deviation value (s), the effect size (d), the pilot trial sample size per treatment group (m) and the number of simulations is set. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s was set to 1, various values of d were investigated between 0.05 and 1, m was chosen to be the optimal value for the selected adjustment method and chosen effect size (from Table 4.8) and 10,000 simulations were carried out.

Step 2: For the pilot trial: Simulate the control arm of sample size m from a Normal distribution with mean 0 and standard deviation s. Simulate the experimental arm of sample size m from a Normal distribution with mean d and standard deviation s. From this pilot data calculate the sample variance.

Step 3: Calculate the sample size required for the main trial (n) based on this pilot data according to the adjustment method selected (80% UCL approach use Equation 4.2 with X=0.8, 95% UCL approach use Equation 4.2 with X=0.95, NCT approach use Equation 2.21).

Step 4: For the main trial: Simulate the control arm of sample size n (from Step 3) from a Normal distribution with mean 0 and standard deviation s. Simulate the experimental arm of sample size n from a Normal distribution with mean d and standard deviation s. Perform a t-test on this data and record the results and the main trial sample size.

Step 5: Repeat steps 2 to 4 10,000 times.

Step 6: Estimate the power of the design by calculating the proportion of trials which rejected the null hypothesis.

Step 7: Calculate the percentage of trials that are larger than the sample size required for 90 and 80% power with a known variance.

For example, if the Type I error rate is chosen to be 0.05 or 5% (two-sided), the Type II error rate is set at 0.1 or 10%, if the chosen standardised effect size is 0.5 and the NCT approach is the chosen adjustment method. The first 5 simulations will give you for example:

Simulation	P-value<0.05	Total Two-armed Sample Size of Main Trial	Greater than 90% power based on the true variance	Greater than 80% power based on the true variance
1	Yes	140	No	Yes
2	Yes	168	No	Yes
3	Yes	108	No	No
4	Yes	198	Yes	Yes
5	Yes	224	Yes	Yes

The algorithm would continue until simulation number 10,000 after which the average main trial sample size, power, and proportion of trials which are large to have 90 or 80% power are calculated.

Table 4.15: Average Power for Two-Armed Trials Designed Using Different Adjustment Methods Based on 10,000 Simulations Using 90% Power, 5% Type I Error Rate and 'Optimal' Pilot Trial Sample Sizes

Standardised Effect Size		80% UCL	95% UCL	NCT
0.05	**Pilot Trial Sample Size**	506	794	212
	Average Power	91.25	92.31	90.52
	Percentage of Trials with Power above 90%	81.71	95.31	57.91
	Percentage of Trials with Power above 80%	100.00	100.00	99.87
0.1	**Pilot Trial Sample Size**	210	332	108
	Average Power	92.23	93.28	90.34
	Percentage of Trials with Power above 90%	82.38	95.53	60.34
	Percentage of Trials with Power above 80%	99.99	100.00	99.00
0.2	**Pilot Trial Sample Size**	90	144	56
	Average Power	93.17	94.75	90.36
	Percentage of Trials with Power above 90%	83.40	95.87	64.2
	Percentage of Trials with Power above 80%	99.68	100.00	96.15
0.5	**Pilot Trial Sample Size**	32	50	24
	Average Power	94.37	96.56	92.09
	Percentage of Trials with Power above 90%	84.19	96.26	68.90
	Percentage of Trials with Power above 80%	97.89	99.85	91.00
0.8	**Pilot Trial Sample Size**	20	32	20
	Average Power	95.37	97.60	92.10
	Percentage of Trials with Power above 90%	84.73	95.85	69.45
	Percentage of Trials with Power above 80%	95.33	99.53	89.23

The 80% and 95% UCL approaches are more conservative than the NCT approach, and their average powers are higher than the 90% nominal level. The NCT approach gives the nominal power (of 90%) in >50% of the trials as can be seen in Table 4.15. A large proportion of the trials have >80% power for all adjustment methods. The NCT approach requires consistently smaller sample sizes than the UCL approaches. The cost (in terms of sample size) of the NCT approach is that it only provides the nominal required power on average whereas the UCL approaches, provides the nominal power in at least $100X\%$ of trials (where an $100X\%$ UCL is used). Therefore, the UCL approaches are more conservative than the NCT approach.

4.8 Stepped Rules of Thumb

In many trials the actual value of the standardised difference to be used in the main trial may not be known before the pilot planning stage. This is one of the reasons that the existing rules of thumb for pilot trial sample sizes are so attractive. However, an investigator may know whether the standardised difference is likely to be small, medium or large within some range.

From the results presented, it would seem that any rule of thumb should be stepped and not flat so that the pilot is larger for smaller standardised effect sizes and smaller for larger standardised effect sizes. Therefore Table 4.8 has been used to formulate new stepped rules of thumb for pilot trial sample sizes; these are presented in Table 4.16. They were carried out by grouping the standardised effect sizes (δ) together into extra small ($\delta < 0.1$), small ($0.1 \leq \delta < 0.3$), medium ($0.3 \leq \delta < 0.7$) and large ($\delta \geq 0.7$) and formulating a sample size roughly applicable to all the effect sizes within the band. These offer standard sample sizes for pilot trials which vary depending on whether the standardised effect size for the main trial will be extra small, small, medium or large, from Cohen's classifications. An additional category of extra small has been inserted which, represents standardised effect sizes of 0.1 or less; this is because the results for these trials were many times larger than for a standardised effect size of 0.2.

For the NCT method with an 90% powered main trial the stepped rules are; for very small standardised differences, use a sample size of 150, for small standardised effect sizes, use 50 participants, for medium standardised effect sizes use a sample size of 30, and for any standardised effect size, use a pilot trial sample size of 20 participants.

It should be noted that if the standardised difference to be used in the main trial is known it is still recommended to use the exact calculation.

The recommended stepped rules are based on using the NCT approach to allow for the variance being an estimate of the population value, it uses the distribution of the variance to ensure that the power is achieved on average and converges more quickly to the variance known case than the UCL approaches studied. The UCL approaches are much more conservative and require consistently more participants than the NCT approach.

Table 4.16: Stepped Rules of thumb for Pilot Trial Sample Size using the NCT Approach for a Two-armed Trial

Standardised Effect Size	80% Powered Main Trial	90% Powered Main Trial
Flat Rules		
Extra Small ($\delta < 0.1$)	100	150
Small ($0.1 \leq \delta < 0.3$)	40	50
Medium ($0.3 \leq \delta < 0.7$)	20	30
Large ($\delta \geq 0.7$)	20	20
Proportional Rules		
Extra Small ($\delta < 0.1$)	1%	1%
Small ($0.1 \leq \delta < 0.3$)	5%	6%
Medium ($0.3 \leq \delta < 0.7$)	18%	15%
Large ($\delta \geq 0.7$)	42%	30%

The distances, shown in Table 4.17, of the stepped rules of thumb from the optimal overall sample size (from Table 4.8) are lower than the distances observed when using the

existing rules of thumb in Table 4.6. For 90% powered trials with a small standardised effect size of 0.2 (using the NCT approach) the existing rules of thumb could lead to an extra 48 participants over the minimum possible overall sample size. Using the stepped rule of thumb this possible over recruitment is reduced to 26 participants. The stepped rules of thumb are very close to the optimal value in the majority of these cases.

Table 4.17: Distances for a Two-armed Trial from Optimal Values for the Stepped Rules of Thumb for Varying Standardised Effect Sizes

Standardised Effect Size	Stepped Rule of Thumb Pilot Trial	Overall Sample Size	Optimal Pilot Trial Sample Size	Optimal Overall Trial Sample Size	Distance Between Pilot Trial Sample Sizes	Distance Between Overall Sample Sizes
90% Powered Main Trial						
0.05	150	17,260	212	17,234	-62	26
0.2	50	1,162	56	1,160	-6	2
0.5	30	216	24	214	+6	2
0.8	20	98	20	98	0	0
80% Powered Main Trial						
0.05	100	12,878	148	12,854	-48	24
0.2	40	862	38	862	+2	0
0.5	20	160	20	160	0	0
0.8	20	76	20	76	0	0

4.9 Summary

The NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC) define pilot trials in context of the planning of a future trial (NETSCC, 2012). Therefore, the method of minimising the sample size across trials could be thought to be the most appropriate as it treats the pilot trial as part of the whole trial programme rather than a stand-alone trial.

The aims of this chapter were to: find the theoretical minimum overall sample size using the NCT and the UCL approaches and hence calculate the optimal pilot trial sample size,

which leads to this minimum overall sample size (Sections 4.2, 4.3 and 4.4), compare the existing rules of thumb (both flat and proportional) to the optimal results (Section 4.5 and 4.6, Table 4.8 and Table 4.9); and the use these results to build new stepped rules of thumb (Section 4.8, Table 4.16), as well as to compare the effect of the NCT and the UCL approaches on the trial power, this was investigated as a simulation study and presented in Section 4.7.

In order to achieve the aims, set out in Section 4.1.1, this chapter proposes a method for estimating the sample size of a pilot trial in order to minimise the overall sample size, i.e. the sample size of the pilot and main trial together, for different adjustment methods. It demonstrated how the size of the pilot trial impacts on the size of the overall trial when either the UCL approach or the NCT method is used to calculate the sample size for the main trial (Figures 4.7 to 4.9).

If the pilot is large the main trial will be relatively small or, if the pilot is small the main trial will be relatively large. It can be seen from the results that the NCT approach provides lower overall sample sizes than any of the other methods (Section 4.4, Table 4.5) while maintaining the average power at the nominal level (Section 4.7, Table 4.15). There are situations in which using one of the more conservative methods would add only a few more patients to the required sample size in such situations an investigator may choose to use the larger sample size to be more sure of achieving the required power and to gain a more accurate estimate of the treatment effect. However, there are circumstances as have been demonstrated in this chapter where using one of the conservative approaches can lead to much larger sample size than would be required to ensure the required power on average. For example, if the anticipated standardised effect size for a trial is 0.5 with a planned power of 90% the NCT approach would lead to a total overall sample size of 214 and the 80% UCL approach would lead to a sample size of 248. Here the increase in sample size is only 34 participants therefore the investigator might want to be more conservative and use the larger sample size. However, for a standardised effect size of 0.3 and 90% power the NCT approach would lead to an overall trial sample size of 542, whereas the

80% UCL approach would lead to a sample size of 616. Here the difference is 74 participants, therefore the investigator might be willing to sacrifice some certainty of achieving the required power to reduce the sample size requirement of the trial.

The results in the chapter show that as the sample size of the main trial increases, the size of the pilot trial should also increase. For medium effect sizes, the existing rules seem sufficient (Figure 4.8); however, as we move away from a standardised effect size of 0.5 the flat rules of thumb can over or underestimate the pilot trial sample size that would minimise the overall sample size (Figures 4.7 and 4.9). It was highlighted how when using a flat rule or a proportional rule of thumb for justifying a sample size choice for an external pilot trial without considering the main trial you could end up using a pilot trial sample size we which will result in a suboptimal sample size for the overall trial when using one of the adjustment methods in the main trial sample size calculation. Therefore, using these flat rules of thumb would lead to more patients than theoretically required being recruited to the overall trial; this is especially seen at small standardised effect sizes.

The results in Section 4.6 investigate the use of proportional rules, which use proportional pilot trial sample sizes to the main trial sample size. It was found that no one proportion is optimal for all standardised effect sizes. Although the proportional methods are an improvement on the flat rules of thumb in that they allow the pilot to be larger for large main trials and smaller for small main trials. They still place a restriction on the sample size of the pilot trial, which means that the optimal overall sample size may not be achieved.

To look at which proportions would be optimal for this kind of rule of thumb Table 4.8 was used to calculate what percentage of the main trial the optimal pilot trial would be, these results can be seen in Table 4.11, therefore restricting the pilot to be a certain percentage has the same effects as using a flat rule of thumb; it cause an over recruitment compared to what is theoretically possible.

The NCT approach to set the main trial sample size in conjunction with the method presented of calculating a pilot trial sample size is recommended. Doing so will on average maintain the nominal power requirement and minimise the overall sample size for the pilot and the main trial together. If simpler calculations are to be carried out for a pilot trial the stepped rules proposed in this chapter are recommended. However, if the standardised effect size to be used in the main trial is known, it is recommended that the procedure outlined in Section 4.4 be used to find the optimal pilot trial sample size for the study rather than using the flat rules of thumb as a guide.

This chapter focussed on external pilot trials. Chapters 6 and 7 will look at the design of internal pilot trials. Minimising the overall sample size does not only have ethical advantages for numbers of patients used, but also for the financial cost of trials. However, depending on the relative costs between the pilot and the main trial minimising the sample size may not necessarily minimise the overall cost of the trial. Hence Chapter 5 investigates using information about the relative cost of the pilot compared to the main trial to minimise the overall financial cost of a trial programme.

Chapter 5

Minimising the Overall Financial Cost of a Trial

5.1 Introduction

In publicly funded research a major concern is to get the best value for money from a trial. The cost of a trial is strongly linked to the number of participants recruited. However, the cost of entering a patient into a pilot trial may not be equal to the cost of recruiting a patient into a main trial. Minimising the number of people within a trial will therefore not necessarily minimise the total financial cost of the overall trial. For example, the CACTUS trial (a description of which can be found in Appendix C) was a pilot trial with two treatment arms and aimed to enrol 30 participants (Palmer et al., 2011). The pilot trial cost £279,000, therefore around £9,300 per participant. The main trial (Big Cactus) had a sample size requirement of 285 with three treatment arms and cost £1.5 million, a cost of £5,264 per participant (Palmer, 2015, CTRU, 2016).

This chapter extends the work on minimising the overall number of participants described in Chapter 4 to include a factor representing the relative cost of entering a patient in to the main trial compared to the pilot, with the aim of minimising the overall financial cost of the trial. The chapter investigates how the balance of sample sizes between the two trials affects the overall cost of both the pilot and the main trial together allowing for uneven costs between trials. Rules for pilot trial sample size to minimise the overall trial cost will be derived based on the results.

5.1.1 Aims

This chapter aims to extend the work presented in Chapter 4 on minimising the overall trial sample size by including the relative cost of a pilot and main trial to:

- Minimise the overall cost of the pilot and the main trial together using the NCT and the UCL methods,
- Find the theoretical 'optimal' values of the overall trial cost which could be achieved using these methods,
- Calculate the pilot trial sample size, which leads to this optimal value,
- Develop new rules of thumb which aim to minimise the overall financial cost of the trial and,
- Compare these new rules of thumb to the ones that minimise the overall trial sample size.

This chapter similarly to Chapter 4 focuses on external pilot trials, which are primarily looking at estimating the variance to be used in the main trial sample size calculation and where there are no changes between the pilot and the main trial, which would affect the generalizability of the variance estimate.

5.2 Minimising the Overall Financial Cost

If an adjustment method is used to calculate the main trial sample size based on the size of the pilot trial, the work in Chapter 4 showed that it is possible to choose the pilot trial sample size in order to minimise the overall trial sample size.

If the cost of a trial is directly related to the sample size of a trial, then it is also possible to minimise the cost of a trial using the same methods. The overall cost of the trial (C) can be expressed through the function,

$$C = C_P M + C_M N_M, \quad (5.1)$$

where, C_P is the cost per participant in the pilot trial, C_M is the cost per participant in the main trial, M is the sample size of the pilot and N_M is the sample size of the main trial calculated using either Equation 4.2 or Equation 2.21 for the UCL or the NCT approach respectively or Equation 2.15 for an unadjusted main trial sample size calculation.

If $C_P/C_M = R$ where R is the relative cost of a participant in the pilot trial to the main trial then,

$$C = RC_M M + C_M N_M,$$

therefore,

$$C/C_M = RM + N_M, \quad (5.2)$$

thus, by fixing R we can minimise the equation to find the minimum overall trial cost divided by the main trial cost per participant which is in turn a function of the overall trial cost. This is an extension of the methods presented in Chapter 4 when the aim was to minimise the overall sample size through Equation 4.1, however, now the aim is the minimise the value of the cost ratio.

To find the sample size, which would lead to the minimum of this equation, the size of the pilot trial is varied over a range of values and the required main trial sample size is calculated, for a given standardised effect size and relative cost ratio.

Once the minimum C/C_M has been found, the pilot trial and main trial sample sizes, which produce this minimum, can be derived. This process is shown in Figure 5.1 for when the NCT approach is used to adjust the main trial sample size calculation and Figure 5.2 for when 80% UCL approach is used. In the previous chapter it was seen that the 95% UCL approach led to very conservative results and therefore only the results for the NCT and the 80% UCL approaches are presented from this point onwards. Although the methods are still applicable to the 95% UCL approach.

Figure 5.1: Process for Finding the Sample Size to Minimise the Overall Trial Cost for the NCT Approach

Step 1: Values for Type I (α) and Type II (β) error are selected, the standard deviation value (s) and the effect size (d) is set, and the starting value of m is chosen. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s was set to 1, various values of d were investigated between 0.05 and 1 and the starting value of m was chosen to be 2 participants per treatment group to prevent the degrees of freedom for the variance from being less than or equal to zero. Set $i = 1$.

Step 2: For a pilot trial sample size m_i, where i is the iteration number, estimate the main trial sample size, n_{START} from Equation 2.22.

Step 3: Using n_{START} as a starting point for n_M in Equation 2.21 iterate n_M upwards until the inequality in Equation 2.21 is satisfied.

Step 4: Estimate C/C_M from Equation 5.2.

Step 5: For $i = 1$ go to Step 6, for $i > 1$ go to Step 7.

Step 6: Add 1 to the previous pilot trial sample size, m_i and i, and go to Step 2.

Step 7: If $C/C_M$for $m_i \leq C/C_M$for m_{i-1} then go to Step 6. If $C/C_M$for $m_i > n_T$ for m_{i-1} then go to Step 8.

Step 8: Take n_M for m_{i-1} as the main trial sample size, which minimises the overall cost of the trial.

Step 9: Take m_{i-1} for m_{OPT} as the pilot trial sample size, which leads to the minimum overall cost for the trial.

For example, if the Type I error rate is chosen to be 0.05 or 5% (two-sided), the Type II error rate is set at 0.1 or 10%, if the chosen standardised effect size is 0.2 and the NCT approach is the chosen adjustment method. The first 5 pilot trial sample sizes will give you the following results for a two armed trial:

Pilot Trial Sample Size	Main Trial Sample Size	Overall Sample Size
4	4412	4416
6	2076	2082
8	1640	1648
10	1464	1474
12	1368	1380

This continues over a range of pilot sample sizes, recording the results each time. After a large number of sample sizes have been investigated the minimum of the function is found and the corresponding pilot trial and main trial sample sizes. In the case presented this method would lead to selecting a pilot trial sample size of 78, a main trial sample size of 1090 and therefore an overall trial sample size 1168 for a two-armed trial.

Figure 5.2: Process for Finding the Sample Sizes, which lead to the Minimum Overall Trial Cost for the UCL Approach

Step 1: Values for Type I (α) and Type II (β) error are selected the standard deviation value (s) and the effect size (d) is set, the required level of X is chosen, and the starting value of m is chosen. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s was set to 1, various values of d were investigated between 0.05 and 1 and the starting value of m was chosen to be 2 participants per treatment group to prevent the degrees of freedom for the variance from being less than or equal to zero. Set $i = 1$.

Step 2: For a pilot trial sample size m_i, where i is the iteration number, estimate the main trial sample size, n_M from Equation 4.2.

Step 3: Estimate C/C_M from Equation 5.2.

Step 4: For $i = 1$ go to Step 5, for $i > 1$ go to Step 6.

Step 5: Add 1 to the previous pilot trial sample size, m_i and i, and go to Step 2.

Step 6: If C/C_M for $m_i \leq C/C_M$ for m_{i-1} then go to Step 5. If C/C_M for $m_i >$ C/C_M for m_{i-1} then go to Step 7.

Step 7: Take n_M for m_{i-1} as the main trial sample size, which minimises the overall cost of the trial.

Step 9: Take m_{i-1} for m_{OPT} as the pilot trial sample size, which leads to the minimum overall cost for the trial.

5.3 Optimal Values of the Pilot and Main Trial Sample Size

Figures 5.3 to 5.6 display the pilot trial sample size plotted against the function of the trial cost C/C_M for the standardised effect sizes 0.05, 0.2, 0.5 and 0.8 respectively, with 90% power for the main trial and using the NCT approach. The lines represent different values of the relative cost of the pilot trial to the main trial. The black solid line represents a relative cost of 0.5 i.e. the main being twice as expensive as the pilot trial per participant. Following on the red dashed line, the green dotted and dashed line, the purple dotted line and the light blue dotted and dashed line represent the relative costs of 1, 2, 10 and 50 respectively.

Figure 5.3: Comparing the Overall Trial Cost for the NCT Approach for Varying Values of Relative Cost and a Standardised Effect Size of 0.05

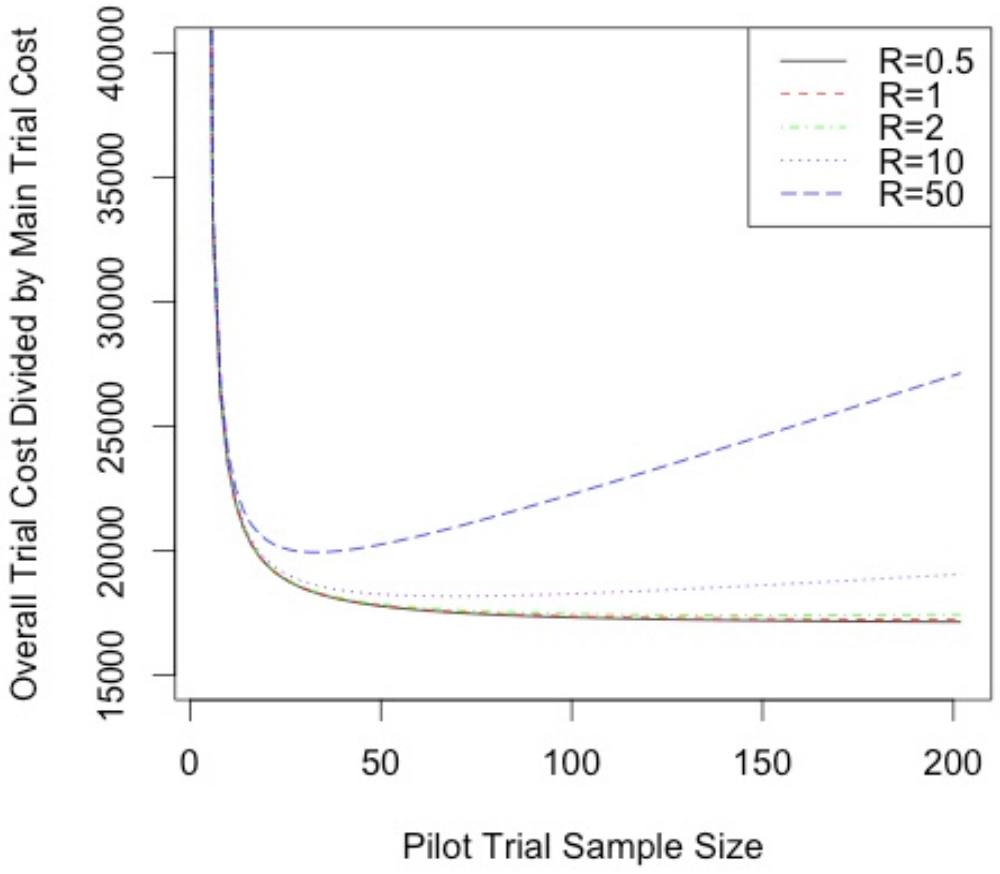


Figure 5.4: Comparing the Overall Trial Cost for the NCT Approach for Varying Values of Relative Cost and a Standardised Effect Size of 0.2

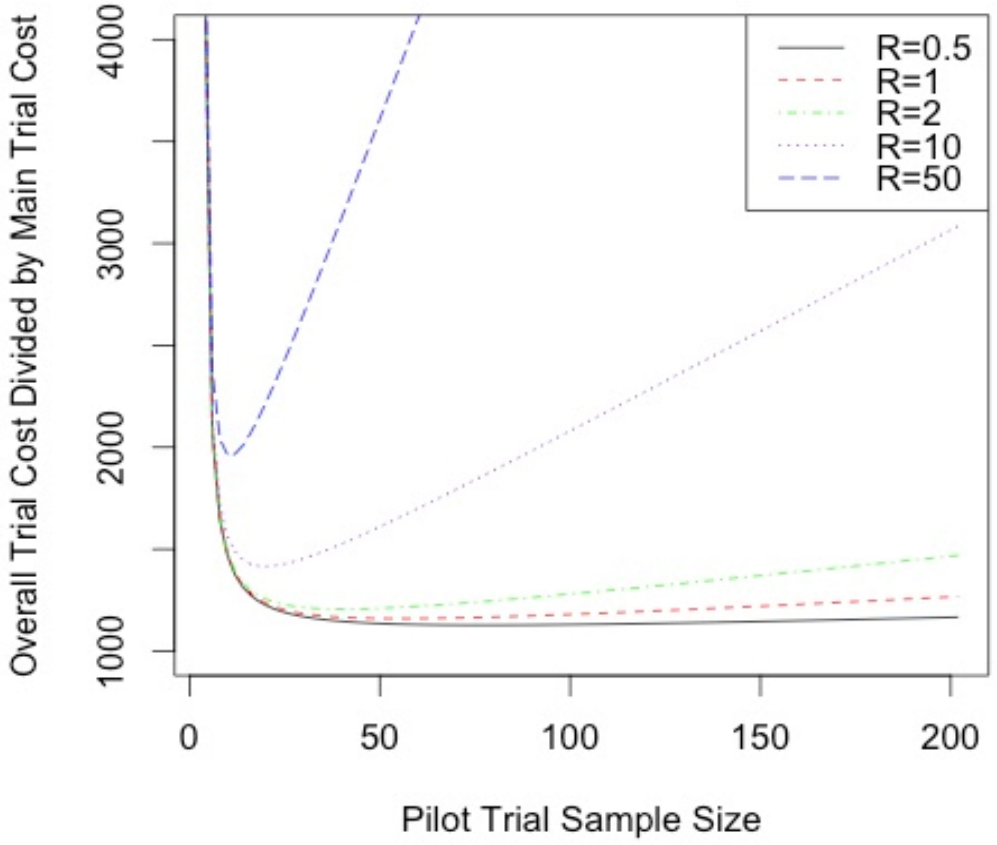


Figure 5.5: Comparing Overall Trial Cost for Varying for the NCT Approach for Varying Values of Relative Cost and a Standardised Effect Size of 0.5

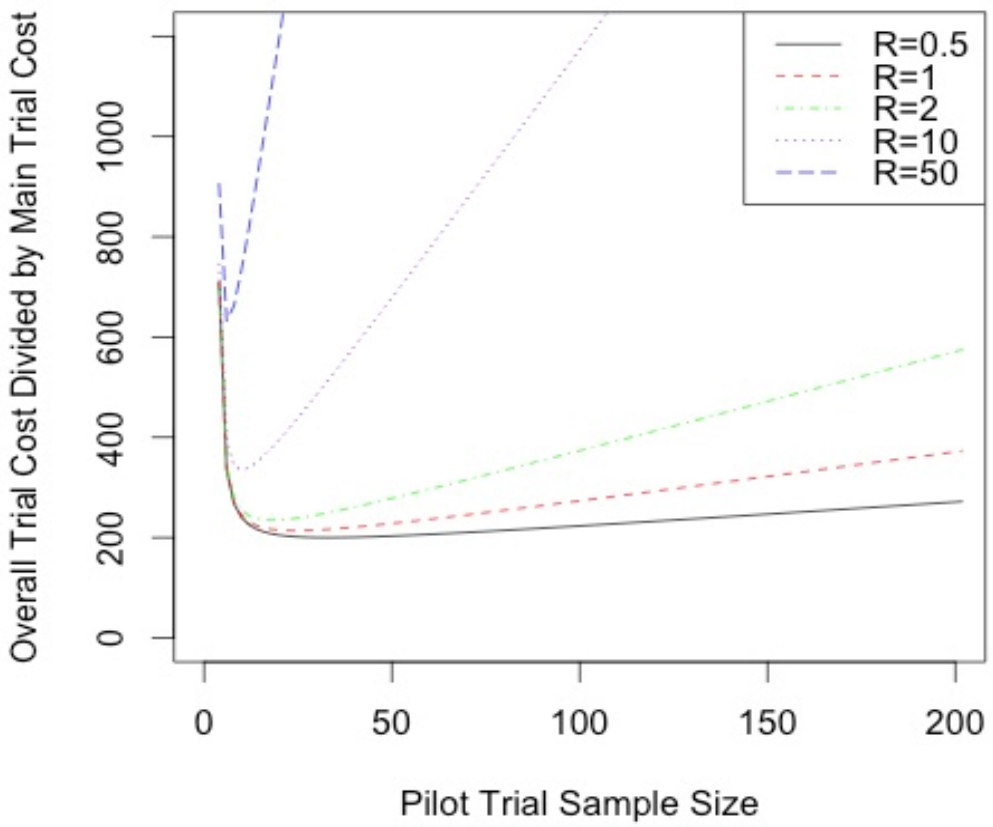
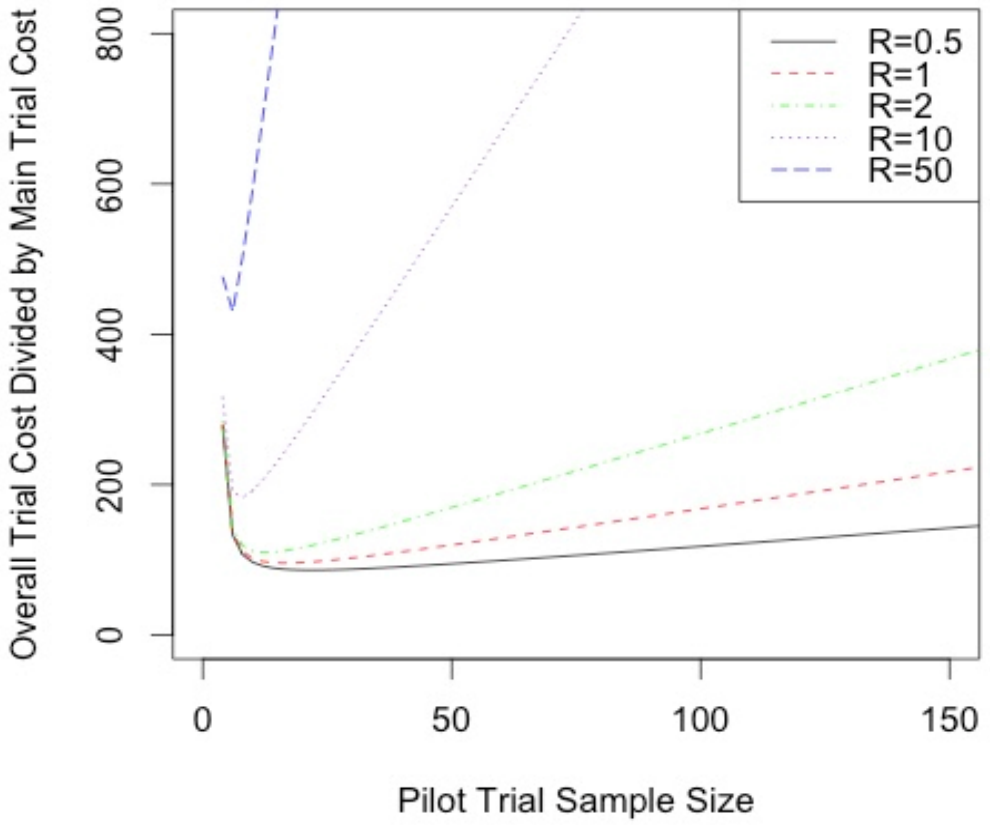


Figure 5.6: Comparing Overall Trial Cost for Varying for the NCT Approach for Varying Values of Relative Cost and a Standardised Effect Size of 0.8



It can be seen that we can find a minimum for C/C_M so that we can derive the pilot trial sample size, which leads to the minimum overall trial cost. As the standardised effect size increases (and so the required main trial sample size decreases) the optimal pilot trial sample size decreases. C/C_M is a function of the overall cost of the trial, it has no real meaning in itself however, by minimising this function we can calculate the sample sizes which would lead to the minimum overall cost for the trial.

Tables 5.1 to 5.4 present the results using both the 80% UCL approach and the NCT method to adjust the main trial sample size based on the pilot trial sample size to look at

how the minimum pilot and minimum overall sample sizes are affected by changing the focus from minimising the number of patients to minimising the cost of the trials.

The results are presented for standardised differences of 0.05, 0.2, 0.5 and 0.8 and also for a range of relative costs, where the relative cost is the factor by which the entering someone into the pilot trial costs more than entering a patient into the main trial. Therefore, a relative cost of less than one would mean that the main trial is more expensive per participant than the pilot trial. Likewise, if the relative cost is greater than one the cost per participant in the pilot trial is higher than in the main trial.

Table 5.1: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.05 for a Two-armed Trial

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	752	17,582	18,334	252	16,988	17,240
1.00	506	17,760	18,266	212	17,022	17,234
1.25	440	17,836	18,276	190	17,046	17,236
1.50	392	17,902	18,294	174	17,070	17,244
2.00	326	18,014	18,340	152	17,108	17,260
2.50	284	18,110	18,394	136	17,144	17,280
3.00	254	18,192	18,446	124	17,174	17,298
5.00	184	18,458	18,642	98	17,278	17,376
10.00	122	18,900	19,022	70	17,476	17,546
15.00	96	19,216	19,312	58	17,624	17,682
20.00	80	19,474	19,554	50	17,746	17,796
25.00	70	19,698	19,768	46	17,858	17,904
30.00	64	19,890	19,954	42	17,968	18,010
40.00	54	20,200	20,254	36	18,146	18,182
50.00	48	20,512	20,560	34	18,280	18,314
100.00	32	21,588	21,620	24	18,922	18,946
80% Powered Main Trial						
0.50	654	13,176	13,830	208	12,664	12,872
1.00	420	13,342	13,762	148	12,706	12,854
1.25	366	13,402	13,768	134	12,724	12,858
1.50	326	13,458	13,784	122	12,738	12,860
2.00	272	13,552	13,824	106	12,766	12,872
2.50	236	13,630	13,866	96	12,790	12,886
3.00	212	13,698	13,910	88	12,812	12,900
5.00	154	13,918	14,072	68	12,886	12,954
10.00	102	14,290	14,392	50	13,020	13,070
15.00	80	14,546	14,626	42	13,116	13,158
20.00	68	14,772	14,840	36	13,200	13,236
25.00	60	14,952	15,012	32	13,290	13,322
30.00	54	15,118	15,172	30	13,344	13,374
40.00	46	15,360	15,406	26	13,480	13,506
50.00	40	15,628	15,668	24	13,568	13,592
100.00	28	16,474	16,502	18	13,970	13,988

Table 5.2: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.2 for a Two-armed Trial

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	138	1,172	1,310	78	1,090	1,168
1.00	92	1,206	1,298	56	1,104	1,160
1.25	80	1,218	1,298	50	1,112	1,162
1.50	72	1,228	1,300	46	1,116	1,162
2.00	62	1,248	1,310	40	1,128	1,168
2.50	54	1,264	1,318	36	1,136	1,172
3.00	48	1,280	1,328	34	1,142	1,176
5.00	38	1,322	1,360	28	1,168	1,196
10.00	26	1,408	1,434	20	1,216	1,236
15.00	22	1,456	1,478	18	1,252	1,270
20.00	18	1,508	1,526	16	1,288	1,304
25.00	16	1,554	1,570	14	1,310	1,324
30.00	16	1,584	1,600	14	1,336	1,350
40.00	14	1,654	1,668	12	1,368	1,380
50.00	12	1,702	1,714	12	1,410	1,422
100.00	10	1,926	1,936	10	1,538	1,548
80% Powered Main Trial						
0.50	116	886	1,002	54	814	868
1.00	78	914	992	40	824	864
1.25	68	924	992	36	828	864
1.50	62	932	994	34	832	866
2.00	52	948	1,000	30	838	868
2.50	46	960	1,006	26	844	870
3.00	42	974	1,016	24	850	874
5.00	32	1,014	1,046	20	870	890
10.00	22	1,078	1,100	16	898	914
15.00	18	1,128	1,146	12	934	946
20.00	16	1,162	1,178	12	952	964
25.00	14	1,208	1,222	10	976	986
30.00	14	1,236	1,250	10	976	986
40.00	12	1,272	1,284	10	1,008	1,018
50.00	12	1,314	1,326	8	1,052	1,180
100.00	8	1,534	1,542	8	1,116	1,124

Table 5.3: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.5 for a Two-armed Trial

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	48	206	254	34	184	218
1.00	32	216	248	24	190	214
1.25	30	220	250	22	194	216
1.50	26	224	250	20	196	216
2.00	24	230	254	18	200	218
2.50	20	236	256	18	202	220
3.00	20	240	260	16	208	224
5.00	16	254	270	14	216	230
10.00	12	282	294	10	236	246
15.00	10	294	304	10	248	258
20.00	10	308	318	8	264	272
25.00	8	330	338	8	264	272
30.00	8	330	338	8	290	298
40.00	8	360	368	8	290	298
50.00	6	408	414	6	334	340
100.00	6	502	508	6	424	430
80% Powered Main Trial						
0.50	40	158	198	24	138	162
1.00	28	166	194	18	142	160
1.25	26	170	196	16	144	160
1.50	24	172	196	16	144	160
2.00	20	176	196	14	148	162
2.50	18	182	200	12	150	162
3.00	16	186	202	12	154	166
5.00	14	198	212	10	158	168
10.00	10	220	230	8	170	178
15.00	10	230	240	8	180	188
20.00	8	246	254	6	198	204
25.00	8	270	278	6	198	204
30.00	8	270	278	6	198	204
40.00	6	306	312	6	230	236
50.00	6	306	312	6	230	236
100.00	6	376	382	4	320	324

Table 5.4: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.8 for a Two-armed Trial

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	30	86	116	22	76	98
1.00	20	92	112	16	80	96
1.25	18	96	114	16	82	98
1.50	18	96	114	14	84	98
2.00	16	100	116	14	84	98
2.50	14	104	118	12	86	98
3.00	12	108	120	12	90	102
5.00	10	116	126	10	98	108
10.00	8	130	138	8	114	122
15.00	8	142	150	8	114	122
20.00	6	160	166	6	132	138
25.00	6	160	166	6	132	138
30.00	6	160	166	6	132	138
40.00	6	198	204	6	166	172
50.00	6	198	204	6	166	172
100.00	4	296	300	6	166	172
80% Powered Main Trial						
0.50	26	66	92	16	56	72
1.00	18	72	90	12	60	72
1.25	16	74	90	12	60	72
1.50	16	74	90	10	62	72
2.00	14	78	92	10	64	74
2.50	12	80	92	10	64	74
3.00	12	84	96	8	68	76
5.00	10	90	100	8	72	80
10.00	8	106	114	6	78	84
15.00	6	120	126	6	90	96
20.00	6	120	126	6	90	96
25.00	6	120	126	6	90	96
30.00	6	148	154	6	90	96
40.00	6	148	154	4	126	130
50.00	6	148	154	4	126	130
100.00	4	220	224	4	126	130

Tables 5.1 to 5.4 show, as in Chapter 4, that as the standardised effect size increases the required overall sample size and the optimal pilot trial sample size decreases. Additionally, as the relative cost of the pilot trial increases the optimal pilot trial sample size decreases as the extra cost of entering someone into the pilot trial starts to outweigh the accuracy gained in the estimate of the variance and the associated reduction in main trial sample size. Furthermore, it can be seen that if the main trial is more expensive than the pilot trial per participant then the optimal sample size overall also increases compared to the situation where costs are assumed equal.

As in Chapter 4 for some larger standardised effect sizes and for higher relative cost values the optimal pilot trial sample size falls below 20. Having such small pilot sample sizes may be unrealistic to achieve the other objectives of a pilot trial. Therefore, the following tables (Tables 5.5 to 5.8) present the results with a lower cap of 20 participants for the pilot trial.

Table 5.5: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.05 with a Lower Cap of 20 Participants for Two-armed Trials

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	752	17,582	18,334	252	16,988	17,240
1.00	506	17,760	18,266	212	17,022	17,234
1.25	440	17,836	18,276	190	17,046	17,236
1.50	392	17,902	18,294	174	17,070	17,244
2.00	326	18,014	18,340	152	17,108	17,260
2.50	284	18,110	18,394	136	17,144	17,280
3.00	254	18,192	18,446	124	17,174	17,298
5.00	184	18,458	18,642	98	17,278	17,376
10.00	122	18,900	19,022	70	17,476	17,546
15.00	96	19,216	19,312	58	17,624	17,682
20.00	80	19,474	19,554	50	17,746	17,796
25.00	70	19,698	19,768	46	17,858	17,904
30.00	64	19,890	19,954	42	17,968	18,010
40.00	54	20,200	20,254	36	18,146	18,182
50.00	48	20,512	20,560	34	18,280	18,314
100.00	32	21,588	21,620	24	18,922	18,946
80% Powered Main Trial						
0.50	654	13,176	13,830	208	12,664	12,872
1.00	420	13,342	13,762	148	12,706	12,854
1.25	366	13,402	13,768	134	12,724	12,858
1.50	326	13,458	13,784	122	12,738	12,860
2.00	272	13,552	13,824	106	12,766	12,872
2.50	236	13,630	13,866	96	12,790	12,886
3.00	212	13,698	13,910	88	12,812	12,900
5.00	154	13,918	14,072	68	12,886	12,954
10.00	102	14,290	14,392	50	13,020	13,070
15.00	80	14,546	14,626	42	13,116	13,158
20.00	68	14,772	14,840	36	13,200	13,236
25.00	60	14,952	15,012	32	13,290	13,322
30.00	54	15,118	15,172	30	13,344	13,374
40.00	46	15,360	15,406	26	13,480	13,506
50.00	40	15,628	15,668	24	13,568	13,592
100.00	28	16,474	16,502	20	13,804	13,824

Table 5.6: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.2 with a Lower Cap of 20 Participants for Two-armed Trials

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	138	1,172	1,310	78	1,090	1,168
1.00	92	1,206	1,298	56	1,104	1,160
1.25	80	1,218	1,298	50	1,112	1,162
1.50	72	1,228	1,300	46	1,116	1,162
2.00	62	1,248	1,310	40	1,128	1,168
2.50	54	1,264	1,318	36	1,136	1,172
3.00	48	1,280	1,328	34	1,142	1,176
5.00	38	1,322	1,360	28	1,168	1,196
10.00	26	1,408	1,434	20	1,216	1,236
15.00	22	1,456	1,478	20	1,216	1,236
20.00	20	1,472	1,492	20	1,216	1,236
25.00	20	1,472	1,492	20	1,216	1,236
30.00	20	1,472	1,492	20	1,216	1,236
40.00	20	1,472	1,492	20	1,216	1,236
50.00	20	1,472	1,492	20	1,216	1,236
100.00	20	1,472	1,492	20	1,216	1,236
80% Powered Main Trial						
0.50	116	886	1,002	54	814	868
1.00	78	914	992	40	824	864
1.25	68	924	992	36	828	864
1.50	62	932	994	34	832	866
2.00	52	948	1,000	30	838	868
2.50	46	960	1,006	26	844	870
3.00	42	974	1,016	24	850	874
5.00	32	1,014	1,046	20	866	886
10.00	22	1,078	1,100	20	866	886
15.00	20	1,100	1,120	20	866	886
20.00	20	1,100	1,120	20	866	886
25.00	20	1,100	1,120	20	866	886
30.00	20	1,100	1,120	20	866	886
40.00	20	1,100	1,120	20	866	886
50.00	20	1,100	1,120	20	866	886
100.00	20	1,100	1,120	20	866	886

Table 5.7: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.5 with a Lower Cap of 20 Participants for Two-armed Trials

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	48	206	254	34	184	218
1.00	32	216	248	24	190	214
1.25	30	220	250	22	194	216
1.50	26	224	250	20	196	216
2.00	24	230	254	20	196	216
2.50	20	236	256	20	196	216
3.00	20	236	256	20	196	216
5.00	20	236	256	20	196	216
10.00	20	236	256	20	196	216
15.00	20	236	256	20	196	216
20.00	20	236	256	20	196	216
25.00	20	236	256	20	196	216
30.00	20	236	256	20	196	216
40.00	20	236	256	20	196	216
50.00	20	236	256	20	196	216
100.00	20	236	256	20	196	216
80% Powered Main Trial						
0.50	40	158	198	24	138	162
1.00	28	166	194	20	140	160
1.25	26	170	196	20	140	160
1.50	24	172	196	20	140	160
2.00	20	176	196	20	140	160
2.50	20	176	196	20	140	160
3.00	20	176	196	20	140	160
5.00	20	176	196	20	140	160
10.00	20	176	196	20	170	160
15.00	20	176	196	20	140	160
20.00	20	176	196	20	140	160
25.00	20	176	196	20	140	160
30.00	20	176	196	20	140	160
40.00	20	176	196	20	140	160
50.00	20	176	196	20	140	160
100.00	20	176	196	20	140	160

Table 5.8: Optimal Pilot Trial Sample Sizes and Minimum Overall Sample Sizes for both Adjustment Methods to Minimise the Overall Trial Cost for a Standardised Effect Size of 0.8 with a Lower Cap of 20 Participants for Two-armed Trials

	80% Upper Confidence Limit Approach			Non-Central T-distribution Approach		
Relative Cost	**Optimal Pilot**	**Main Trial**	**Overall Trial**	**Optimal Pilot**	**Main Trial**	**Overall Trial**
90% Powered Main Trial						
0.50	30	86	116	22	76	98
1.00	20	92	112	20	78	98
1.25	20	92	112	20	78	98
1.50	20	92	112	20	78	98
2.00	20	92	112	20	78	98
2.50	20	92	112	20	78	98
3.00	20	92	112	20	78	98
5.00	20	92	112	20	78	98
10.00	20	92	112	20	78	98
15.00	20	92	112	20	78	98
20.00	20	92	112	20	78	98
25.00	20	92	112	20	78	98
30.00	20	92	112	20	78	98
40.00	20	92	112	20	78	98
50.00	20	92	112	20	78	98
100.00	20	92	112	20	78	98
80% Powered Main Trial						
0.50	26	66	92	20	56	76
1.00	20	70	90	20	56	76
1.25	20	70	90	20	56	76
1.50	20	70	90	20	56	76
2.00	20	70	90	20	56	76
2.50	20	70	90	20	56	76
3.00	20	70	90	20	56	76
5.00	20	70	90	20	56	76
10.00	20	70	90	20	56	76
15.00	20	70	90	20	56	76
20.00	20	70	90	20	56	76
25.00	20	70	90	20	56	76
30.00	20	70	90	20	56	76
40.00	20	70	90	20	56	76
50.00	20	70	90	20	56	76
100.00	20	70	90	20	56	76

It can be seen that the pilot sample sizes become the imposed minimum of 20 quicker for larger effect sizes and when using the NCT approach over the 80% UCL approach. The table below displays the smallest relative cost for which the NCT approach would result in a pilot trial sample size of 20.

Table 5.9: The Smallest Relative Cost for which the NCT Approach leads to a Pilot Trial Sample Size of 20

Standardised Effect Size	80% Powered Main Trial	90% Powered Main Trial
0.05	100.0	>100.0
0.2	5.0	10.0
0.5	1.0	1.5
0.8	<0.5	1.0

5.4 Rules of Thumb

In Chapter 4 rules of thumb were derived that summarised the calculations, which aimed to minimise the number of participants in the overall trial, of the combined pilot and main trial (Table 4.16). In this section the rules of thumb presented aim to derive approximate required pilot trial sample sizes if the aim of the sample size justification is to minimise the overall cost of the pilot and the main trial. These rules presented in Table 5.10 allow for the cost of the pilot and the main trial per participant to vary as well as the standardised effect size.

Table 5.10: Rules for Thumb for Pilot Trial Sample Size for a Two-armed Trial to Minimise Overall Trial Cost

Standardised Effect Size	Relative Cost	80% Powered Main Trial	90% Powered Main Trial
Extra Small	**R < 1**	240	260
	R = 1	100	150
	1 < R ≤ 5	90	140
	5 < R ≤ 20	50	60
	R > 20	30	40
Small	**R < 1**	60	80
	R = 1	40	50
	1 < R ≤ 5	30	40
	5 < R ≤ 20	20	20
	R > 20	20	20
Medium	**R < 1**	30	40
	R = 1	20	30
	1 < R ≤ 5	20	20
	5 < R ≤ 20	20	20
	R > 20	20	20
Large	**R < 1**	20	30
	R = 1	20	20
	1 < R ≤ 5	20	20
	5 < R ≤ 20	20	20
	R > 20	20	20

As may have been expected when the relative cost of the pilot to the main moves away from one the optimal sample sizes change. If the pilot is to be less expensive per participant than the main trial the rules say to include more people in the pilot trial, increasing accuracy of the estimates and thus reducing the required sample size in the relatively expensive main trial. If the pilot is to be more expensive than the main trial per participant then the rules suggest including fewer people in the relatively expensive pilot trial, thus reducing the accuracy of the estimate of the variance and increasing the required sample size of the main trial. The more expensive the pilot trial compared to the main trial the less people are included in the pilot trial and the bigger the main trial will be. These rules to minimise costs require more participants overall than if we were looking to minimise the number of participants involved in the overall trial that are required to estimate the variance sufficiently well. This effect has more impact on the smaller effect sizes.

Please note that for R = 1 the rules of thumb in Table 5.10 are the same as the stepped rules given in Table 4.16 in Chapter 4.

5.5 Summary

In Chapter 4 work was presented, which aimed to minimise the overall number of participants used in a trial. Although this is important for ethical reasons, it also has cost implications for the trial. If the cost per participant is not equal between the pilot and the main trial, using the optimal results as presented in Chapter 4 to minimise the overall number of participants may not minimise the overall cost of your trial.

Minimising the cost of your trial may be an important consideration particularly if the trial is publicly funded. Minimising the cost of the overall trial may mean we use a suboptimal sample size in terms of the number of participants involved across the trials. When applying for funding for a publicly funded trial the investigator must estimate how much the trial will cost to run. This has to be done early in the development of the trial and therefore it is likely that most investigators would be able to have a good idea or could estimate if required the cost of their planned trial. The AcoRD website gives advice on costing research in the NHS (http://www.amrc.org.uk/our-work/funding-clinical-studies/acord-costing-research-in-the-nhs).

The aims of this chapter were to: calculate the values of pilot trial sample size which lead to the minimum overall trial cost (Section 5.3); develop new rules of thumb for these results (Table 5.10) and to compare to the rules laid out in Chapter 4 (Section 5.4).

In order to achieve these aims the function presented in Equation 5.2 was minimised using the processes described for both the UCL and the NCT approach in Figures 5.1 and 5.2. The optimal pilot trial sample size, which leads to this minimum, was then calculated and the results are presented in Tables 5.1 to 5.4. As in Chapter 4 these results were then

condensed into stepped rules of thumb. However, in this chapter there is the extra factor of relative cost per participant between the pilot and the main trial.

The results show that as the relative cost of the pilot trial increases compared to the main trial the pilot sample size, which would lead to the minimum cost decreases. So that it is better in terms of cost to reduce the number of people in the pilot trial, accept the lower accuracy in the prediction of the variance and increase the main trial sample size accordingly. This may not be possible in circumstances were patients are rare or difficult to recruit, in which case the investigators may be less concerned about the cost of the trial and more concerned about the numbers of participants required.

Moreover, this section of work assumes that the cost of a trial is driven entirely by the numbers of patients recruited this is quite simplistic. In reality there is likely to be a certain level of fixed cost for both the pilot trial and the main trial. It is these fixed costs, which may drive up the cost per patient in the pilot trial. It is likely that most of the time the pilot would be more expensive per participant than the main trial due to the fixed costs involved in running a study. If we assume for the purpose of this that a CTU will be used to help run the trial. When a CTU is used there will be a certain amount of fixed costs involved in running the study e.g. for trial staff and database, these will be the same regardless of the size of the trial. In which case the cost per patient is likely to be higher for a pilot trial compared to a main trial. Nevertheless, there are certain situations where the costs for the main are likely to be higher than the cost per patient in the pilot. For example, if the pilot is conducted in one centre with the help of some students. Therefore, I have considered a situation also where the cost per patient in the main is higher than in the pilot trial.

This chapter as well as Chapters 3 and 4 have investigated sample size justifications for external pilot trials. Where the pilot is conducted as a separate trial from the main trial and analysed separately as a standalone trial. The remaining chapters, 6 and 7 will

investigate the use of and the sample size requirements of internal pilot trials, where the sample comes from within the main trial and will be analysed as part of the main trial.

Chapter 6

Internal Pilot Trials and Sample Size Recalculations

6.1 Introduction

Clinical trials have a high chance of negative results, however, clinical trials are expensive and large trials which end in negative results either through lack of treatment effect or bad trial planning and design can be very costly for the investigators. As discussed in Chapter 2 it is important to have an adequate sample size in a clinical trial for ethical reasons and budgeting purposes, to guarantee sufficient power for a statistical test and to accurately estimate the treatment effect (Friede and Kieser, 2006). Any underestimation of the population variance of the outcome measure in the planning phase of the trial could lead to considerable loss of power in the assessment of treatment effect (Posch et al., 2003). In order to address these problems investigators may plan an adaptive design.

An adaptive design allows the investigators to use accumulating data to modify the trial without undermining the validity and integrity of the trial (Chuang-Stein et al., 2006). There are several adaptations, which could be carried out as part of an adaptive design, including: assessing for futility, assessing for superiority and sample size recalculation or any combination of these. Adaptive designs offer a high amount of flexibility to the investigators and it could be argued they are more ethical as they allow a trial to be stopped early if it is shown that a treatment is inferior or superior so that no more participants than needed are recruited.

One of the more common types of adaption in publicly funded trials is a sample size recalculation (SSR). In a survey of 30 UK clinical trials units, 7 responded saying that they implemented SSR in confirmatory trials and of these the majority (n=4) did so in a restricted blinded manner (Dimairo et al., 2015).

The concentration in this thesis is the adapting of the trial design with a sample size recalculation. A sample size recalculation allows estimation of parameters for a sample size recalculation, from within the main trial but without a hypothesis test taking place (Friede and Kieser, 2006). The recalculation could be implemented by using an internal pilot trial as part of the main trial design. An internal pilot trial is a mechanism employed to gain more accurate predictions of parameters for the sample size calculation amongst other things. An internal pilot with an interim analysis design would additionally involve an assessment of the treatment effect via hypothesis testing part way through the main trial, to offer the chance of early stopping (Friede and Kieser, 2006). If this type of assessment is required at the interim then the Type I error rate should be controlled using either a group-sequential design (Jennison and Turnbull, 1999, Pocock, 1977, O'Brien and Fleming, 1979, Whitehead, 1997) or design based on combining P-values (Bauer and Kohne, 1994).

This chapter will expand on this brief definition and outline how the methodology is implemented with the implications for a trials design and the sample size requirements for the internal pilot trial.

Section 6.2 describes further internal pilot trials and their purpose. Section 6.3 outlines the development of internal pilot trial methods and details how the sample size re-calculation is carried out as part of a clinical trial. Section 6.4 reviews the existing different ideas around how to specify the size of an internal pilot trial. Section 6.5 outlines the limitations of these methods and Section 6.6 summarises the work presented in this chapter and how this is developed further in Chapter 7.

6.1.1 Aims

This chapter aims to:

- Define the meaning of the term internal pilot trial,
- Discuss the procedure for a sample size recalculation,
- Outline the difference between a restricted and unrestricted design,
- Discuss the relative strengths and weaknesses of blinded versus unblinded variance re-estimation,
- Identify existing methods of choosing a sample size for an internal pilot trial,
- Discuss the current approaches and outline the areas of work for Chapter 7.

6.2 Internal Pilot Trial

As defined in Chapter 1, a pilot trial is a study that is: a smaller version of the main trial, focusing on testing trial processes rather than on treatment efficacy, has an intention for further work and guides future sample size calculations. A pilot trials aim is to ensure that the main trial delivers maximum benefit as highlighted in work published during this PhD (Whitehead et al., 2014).

An internal pilot has the same objectives but as opposed to an external pilot trial as discussed in previous chapters, the internal pilot sample forms the first part of the main RCT. NIHR define an internal pilot as the first phase of a substantive study where the data contribute to the final analysis (NETSCC, 2012).

The objective of an internal pilot, which this thesis will be focusing on, is the situation where internal pilot allows the adaptation of the trial by allowing the re-calculation of the required sample size in order to protect from trial under-powering. Like an external pilot an internal pilot trial helps to alleviate the uncertainty involved when estimating the parameters for a sample size calculation. To alleviate this uncertainty an initial proportion

of the main trial is used to estimate the parameter(s) of interest (Wittes and Brittain, 1990).

The parameters of interest for continuous outcome variables include the variance of the outcome measure and the treatment effect. The variance is considered a 'nuisance' parameter that must be estimated however, the treatment effect is of direct interest and the value used in the original calculation will usually have been specified (i.e. as the minimum clinically important difference) rather than estimated (Gallo et al., 2006, Proschan et al., 2003).

There are methods for using the observed treatment effect in the sample size recalculation both to re-estimate the variance and also for a re-assessment of the effect size. Recalculating the required sample size based on the observed treatment effect requires the unblinding of trial data midcourse (Proschan et al., 2003), these methods are discussed further in Section 6.3.3. This thesis will concentrate on sample size recalculations where the variance estimate is re-estimated blind to the treatment effect such that the treatment effect is the same as the pre-specified MCID level used in the sample size calculation.

An internal pilot therefore allows an investigator to re-estimate the variance part way through the trial from the actual trial population; which is an advantage over using historical or external pilot trial data. Another advantage of an internal pilot is that unlike for an external pilot trial the observations taken are included in the final analysis (Wittes and Brittain, 1990).

6.3 Sample Size Recalculation

The technique now known as an internal pilot trial was described by Stein in 1945. His idea was to use an initial proportion of the samples from the main trial to calculate an estimate of the variance from within the trial. After an initial proportion of the responses have been collected a new estimate of the required sample size is calculated using the variance estimate from this sample (N_{RECALC}). The rationale being that this estimate should be closer to the true study variance than an estimate from an external pilot or historical data would be. Any additional responses after the internal pilot are then collected and a hypothesis test carried out. However, in order to maintain the nominal Type I error rate exactly Stein excluded the data after the pilot phase from the variance estimate to be used in the hypothesis test.

Zucker et al. (1999) highlighted that this is an inefficient use of data as observations on the set of patients after the internal pilot are not fully utilised. The Stein approach is therefore not generally used in trial designs. The procedure was labelled an internal pilot study by Wittes and Brittain in 1990. They took the method proposed by Stein and adapted the technique to allow the inclusion of all the data points in the calculation of the final variance estimate.

6.3.1 The Restricted and Unrestricted Design

Wittes and Brittain require that at least the original sample size requirement (N_0) is reached, that is, the sample size needed can only be increased above the pre-specified level as a result of the sample size re-calculation. This is referred to as the restricted sample size re-calculation approach and is the most common method applied in publicly funded clinical trials (Dimairo et al., 2015). The new sample size requirement is taken to be,

$$N_1 = \max(N_0, N_{RECALC}), \tag{6.1}$$

so that the final sample size (N_1) is always at least the sample size stated in the initial sample size calculation (Wittes and Brittain, 1990). The inclusion of the data from both phases of the trial in the final variance estimate assumes that the data in the two phases are independent when in fact they are not, this causes an inflation of the Type I error rate (Wittes and Brittain, 1990). However, simulations (Wittes and Brittain, 1990, Birkett and Day, 1994) have shown that the increase is very small in most cases; between 0 and 0.002 for true variance to projected variance estimate ratios of between 0.5 and 2.

Birkett and Day (1994) showed that the rule restricting the sample size to not be reduced lower than the pre-planned sample size calculation level can lead to a higher than required power and hence an overly large sample size if the original estimate of the variance is too high. Therefore, to tackle this problem of increased power they proposed not carrying out a sample size calculation upfront but instead only specifying a sample size for the internal pilot (M_{INT}). Then the rule of Wittes and Brittain could still then be applied in that the recalculated sample size would be the maximum of the internal pilot trial sample size and the recalculated sample size,

$$\mathrm{N}_1 = \max(M_{INT},\ \mathrm{N}_{\mathrm{RECALC}}). \qquad (6.2)$$

Using this rule, the probability of overpowering the trial is lower. However, they recognise that in practice it is impractical to have no estimate of the final sample size of the trial and perhaps would not be acceptable from a regulators or funders perspective (Birkett and Day, 1994). Additionally, the recommendation that the sample size should not be reduced allows an investigator to avoid problems with interpretation of results later if the difference is not found to be statistically significant but clinically relevant due to an overly optimistic interim estimate of treatment effect or variance (Wittes and Brittain, 1990, Proschan et al., 2003). It has also been highlighted that the demonstration of effect with respect to one outcome is rarely the only objective of a clinical trial (Gould and Shih, 1992) so reducing the sample size would impact on the other areas of investigation.

The restricted approach also helps to maintain the Type I error level. If the sample size is allowed to be reduced then the following situations could arise: if a small variance is observed by chance in the internal pilot phase, the overall trial sample size would be reduced hence increasing the importance of this chance low; if a large variance is observed by chance in the internal pilot phase, the overall trial sample size would be increased hence reducing the importance of this chance high; therefore the estimate of the variance is biased downwards inflating the Type I error rate (Julious, 2009).

With the sample size recalculation approach described above there is potential for the sample size to be increased for trials where if the data were unblinded the treatment might not be sufficiently 'promising' to necessitate the increased sample size. For this reason, some authors suggest only inflating the sample size in the cases where the results are 'promising'. This assessment is based on calculating the conditional power at the interim.

Conditional power is the probability that the test statistic at the end of the trial will be greater than the critical value for the test so that the null hypothesis will be rejected, given the observed data that has been collected up to the interim analysis (Posch et al., 2003). This method however, requires an estimate of the treatment effect to be made at the interim and hence the data is unblinded (Mehta and Pocock, 2011). If the intervention is doing better compared to what was expected then the conditional power will be high, if the conditional power is small then the intervention is not performing as well as was predicted (Friedman et al., 2010).

This approach is called the promising zone method, and the sample size is only inflated if the conditional power is within a predefined promising zone as described in the example below (Mehta and Pocock, 2011). Although rarely implemented in publicly funded trials the promising zone method is frequently used in the private sector (Dimairo et al., 2015). An example of implementing the promising zone approach could be: if the conditional power is ≥ 90% the decision could be to leave the sample size unchanged; whilst if the

conditional power is < 50% the trial would continue as planned or be stopped for futility. The sample size would only increase if the conditional power is ≥ 50% but < 90% so that the conditional power would be equal to 90%.

Due to the early assessment of the treatment effect this procedure can inflate the Type I error rate and the analysis should be adjusted accordingly (Friedman et al., 2010). This approach involves the handling of unblinded interim data, and the requirement to preserve the integrity of the trial suggests that the interim analysis should be carried out by someone independent to the study such as the statistician on the data monitoring committee. This would require a greater level of practical organisation when compared to a fixed sample size trial (Mehta and Pocock, 2011). Using the promising zone approach assumes that the observed effect seen at the interim is the true treatment effect. Moreover, it could be seen as extending a trial in order to acquire a significant P-value, as the sample size is adjusted based on the observed data. This could mean that at the end of the trial the treatment effect on which the trial is powered is too small to be of clinical relevance (Friedman et al., 2010).

6.3.3 Blinded and Unblinded Variance Estimation

For the above recalculation procedures the pooled variance estimator is used (Friede and Kieser, 2006). This estimator requires that the data be unblinded mid-trial, this can reveal information on the treatment effect size which could cause bias (Friede and Kieser, 2006). Thus, if possible it would be desirable to use a procedure where the data can remain blinded. This is the preferred approach from a regulatory point of view, ICH E9 (1998), Section 4.4 (Page 19) states that; 'An interim check conducted on the blinded data may reveal that overall response variances, event rates or survival experience are not as anticipated'. They suggest that because of this a revised sample size calculation could be carried out, however it is stated that; 'The steps taken to preserve blindness and the consequences, if any, for the Type I error and the width of confidence intervals should be explained' *(ICH, 1998, p.19)*

The European Medicine Agency's (EMEA) Committee for Medicinal Products for Human Use (CHMP) add to this by stating:

> Whenever possible, methods for blinded sample size reassessment that properly control the Type I error should be used, especially if the sole aim of the interim analysis is the re-calculation of sample size. *(EMEA, 2007, p.6)*

The within group variance can be estimated by the total variance ignoring the treatment allocation; this method had been shown to work well as long as the treatment effect is not large (Friede and Kieser, 2006). Zucker et al. (1999) however suggested an adjustment to the total variance ($S^2_{1,\text{total}}$) when using blinded data which would make the estimate of the within group variance unbiased if the treatment effect is equal to the level set in the alternative hypothesis (d) usually the MCID,

$$\mathrm{s}^2_{1,\text{adj}} = \mathrm{s}^2_{1,\text{total}} - \frac{M_{INT}}{4(M_{INT}-1)}\, d^2, \tag{6.3}$$

where M_{INT} is the sample size of the internal pilot trial. Equation 6.4 can estimate this, if the internal pilot trial is sufficiently large,

$$\mathrm{s}^2_{1,\text{adj}} = \mathrm{s}^2_{1,\text{total}} - \left(\frac{d}{2}\right)^2. \tag{6.4}$$

6.4 Sample Size for an Internal Pilot Trial

The choice of sample size for an internal pilot trial is an important consideration. An investigator must balance the practical need to conduct the sample size recalculation early in the study with the need to include enough people to get an accurate estimate of the variance for the sample size recalculation (Wittes et al., 1999).

For Stein's procedure (1945) for a single sample, Seelbinder (1953) proposed a method to minimise the expected total sample size of the internal pilot trial and the main trial together (Moshman, 1958). Moshman then extended the method of Seelbinder by placing a bound on the chance of requiring an overly large sample size. By trying to strike a balance between minimising the 95^{th} percentage point of the total sample size (internal pilot trial and the main trial) and the expectation of the total sample size, to select an internal pilot sample size.

Wittes and Brittain (1990) propose to use a sample size of half the initial sample size calculation. Nevertheless, Wittes and Brittain stated that they were interested in what sample size should be used for the internal pilot trial. Birkett and Day (1994) conducted simulations which showed that the expected sample size of the overall trial decreased by only small amounts after the internal pilot trial sample size exceeded at least 10 per arm for a two-armed design (Birkett and Day, 1994). They suggest however, that this is a minimum number and in general the larger the trial the larger the internal pilot trial should be (Sandvik et al., 1996) although there may be little to be gained from using more than 30-40 degrees of freedom to estimate the variance (Birkett and Day, 1994). Wittes et al. (1999) state that choosing a proportion of the study of between 0.25 and 0.75 for the internal pilot trial may be considered practical as a compromise between keeping the sample size small and including enough data for an accurate re-estimation.

Both of these methods fail to take into account the amount of information on which the prior estimate of the variance is based (Sandvik et al., 1996). Sandvik et al. (1996)

proposes a method, which takes this into account when calculating the required size of the internal pilot trial. They stipulate that the size of the internal pilot trial, M_{INT} is proportional to the sample size of the main trial, N_M by some constant A; that is $\mathrm{M}_{INT} = \mathrm{AN}_M$. In addition, M_0 is the amount of information on which the prior estimate of the variance is based and they assume that the outcome variable is Normally distributed. If s_0^2 represents the prior estimate of the variance and σ^2 the true population variance and if $N = \mathrm{gs}_0^2$ where g is some constant then a criterion,

$$\mathrm{Prob}[\mathrm{M}_{INT} > \mathrm{g}\sigma^2] < Q,$$

where $\mathrm{g}\sigma^2$ represents the true sample size required at the true value of the variance. Letting $\mathrm{M}_{INT} = \mathrm{Ags}_0^2$ and noting that $(M_0 - 1)\,\mathrm{s}_0^2/\sigma^2$ follows a χ^2 distribution with $(M_0 - 1)$ degrees of freedom. It follows that,

$$\mathrm{Prob}[(M_0 - 1)\,\mathrm{s}_0^2/\sigma^2 > (M_0 - 1)/\mathrm{A})] < \mathrm{Q},$$

makes the internal pilot trial as large as possible but with the chance of the size of the internal pilot trial exceeding the actual required sample size based on the population standard deviation bounded by some investigator specified probability Q. However, they agree with Birkett and Day (1994) that the internal pilot trial should never be less than 20 participants. Although this is an interesting concept it is not investigated further in this thesis. The paper in which this method is discussed makes no suggestion as to what level of Q might be acceptable and the PhD does not answer this question per se. However, intuitively you would like to keep this small. Related to this question though in the proceeding chapter the thesis investigates sample sizes for internal pilot trials to obtain the best estimates of the sample size for the main trial.

Singer (1999) comments that the method of Sandvik et al (1996) could be improved by considering a problem first raised by Birkett and Day (1994). That is stopping recruitment when M_{INT} (internal pilot trial size) is reached until the sample size recalculation has been

done is considered impractical and may cost the study participants, however, continuing recruitment increases the risk of exceeding the true required sample size (Birkett and Day, 1994). Singer's method offers a solution to this problem, allowing recruitment to continue without the risk of exceeding the true required sample size increasing. Say the recruitment rate is z patients per day and the follow up time in the study is t then, zt patients could be recruited during the waiting time caused by the internal pilot sample size re-estimation. Decreasing the internal pilot sample size by this factor, zt would keep the probability of exceeding the true required sample size at the required level, Q. Although this is a valid point raised by Singer this method is not discussed further in this work.

6.5 Summary

This chapter reviewed the literature on internal pilot trials and conducting a sample size recalculation. An internal pilot trial is a two-stage procedure with no interim hypothesis testing, but which allows for a sample size recalculation based on estimating the nuisance parameter (the variance) for the sample size calculation from the first stage data (Kairalla et al., 2012). This should be done where possible in a blinded manner, as is preferred by regulatory agencies (EMEA, 2007, ICH, 1998). It is possible to re-evaluate the sample size in a restricted (Wittes and Brittain, 1990) or unrestricted (Birkett and Day, 1994) approach.

Owing to the fact that Wittes and Brittain (1990) procedure is restricted to only re-evaluating the sample size upwards; it can produce power and sample sizes much higher than really necessary, when the re-estimated variance is less than the original estimate (Birkett and Day, 1994). In this situation the trial carries on to the original calculated sample size. The achieved power will be greater than the nominal level and an unnecessarily large sample size will have been used (Birkett and Day, 1994). In order to try to alleviate this problem, Birkett and Day proposed an extreme method where no sample size calculation is conducted prior to the start of the trial; only the size of the internal pilot trial is specified. This method would prevent unnecessarily large sample

sizes being used when the prior estimate of the variance is greater than the re-estimated variance, provided that the size of the internal pilot trial is less than the re-estimated sample size (Birkett and Day, 1994). In practice however, this is impractical and some estimate of the study sample size would be required for planning and budgetary purposes.

For these reasons the work in Chapter 7 focuses on looking at the blinded restricted procedure; investigating the implications of the Wittes and Brittain (1990) procedure in terms of trial power and sample size. Furthermore, in terms of investigating the required sample size for an internal pilot trial the intention is to consider the idea of Seelbinder (1953) to minimise the size of the overall study (internal pilot trial and main trial) but for two groups.

Sample size re-estimation suffers from the same disadvantages as the original power analysis prior to the conduct of the trial in that the estimates of the trial parameters are treated as the true values when they are in fact estimated based on the interim data (Chuang-Stein et al., 2006). It is however, possible to extend the same adjustment methods to these estimates as used in Chapters 3, 4 and 5 when considering external pilot trials. The effect of using the adjustment methods on the variance estimate from an internal pilot trial is examined in Chapter 7.

If conducting an internal pilot trial design, the original sample size calculation will be based on a traditional formula with the variance assumed to be known. In practice the estimate may come from an external pilot trial therefore, the original main trial sample size calculation is also based on a sample estimate of the variance. The effect on the power and the optimal sample sizes of a trial, if both an external and internal pilot trial is employed in the trial design will also be explored in Chapter 7.

Chapter 7

The Effect of an Internal Pilot Trial on the Power and Sample Size of a Trial

7.1 Introduction

A pilot trial is a trial which mimics the design of the main trial but does not aim to prove superiority of one treatment over the other, as discussed in Chapter 1, although it may have many other objectives including; checking the feasibility of the larger trial, testing trial procedures and estimating parameters for the sample size calculation (Lancaster et al., 2004, Thabane et al., 2010). In previous chapters' pilot trials conducted external to the main trial were discussed. An internal pilot trial has many of the same objectives of an external pilot trial but, as highlighted in Chapter 6, uses an initial sample of the main trial data, which will then still be included in the final analysis. In this chapter it is considered how an internal pilot trial could be used to inform the main trial sample size through a sample size recalculation

A sample size recalculation uses an initial portion of the main trial participants to re-estimate the variance of the outcome measure used in the initial sample size calculation. This re-estimated variance is then used to recalculate the required sample size for the main trial (Wittes and Brittain, 1990).

Under a restricted procedure if this recalculated sample size is larger than the original the sample size will be increased to this new level. If the recalculated sample size is less than

the original sample size the trial will continue to the originally planned sample size (Wittes and Brittain, 1990).

This chapter examines the effect of an internal pilot trial design on the power level of a trial and the required sample size. The effects of the adjustment methods described in Chapter 2, the UCL approach and the NCT method, are extended to internal pilots.

There are recommendations on the internal pilot trial sample size (Wittes and Brittain, 1990, Birkett and Day, 1994, Wittes et al., 1999) (shown in Table 7.1, where N_0 is the initial sample size estimate at the start of the trial). However, none of these recommendations consider the approach investigated throughout this thesis of minimising the overall trial sample size of the pilot and the main trial together for a two-armed design. In this context this chapter will examine the optimal time point in the trial to carry out an internal pilot trial sample size to minimise the sample size of the main trial.

Table 7.1: Sample Size Recommendations for Internal Pilot Trials for a Two-armed Trial

Author	Recommended Sample Size
Wittes and Brittain (1990)	$0.5N_0$
Birkett and Day (1994)	20
Wittes et al. (1999)	$0.25N_0$ to $0.75N_0$

7.1.1 Aims

This chapter aims to:

- Investigate the effect of undertaking an internal pilot trial on the expected power of the main trial;
- Examine the effect on the main trial power of using the adjustment methods (i.e. the UCL and NCT approaches) at the sample size recalculation;
- Make sample size recommendations for an internal pilot trial;

- Explore methods of adjusting the nominal power and power at the sample size recalculation to optimise the overall expected power of the trial and,
- Investigate the effect of assuming the initial value of the variance in the original sample size calculation to also be a sample estimate.

7.2 The Power of a Trial When Using an Internal Pilot Trial – Assuming the Variance is known

This section investigates how using a sample size recalculation procedure at the end of an internal pilot trial affects the power of the whole main trial.

If the variance estimate is assumed known in the initial sample size, N_0, the initial sample size requirement is calculated from Equation 7.1 (also Equation 2.15),

$$N_0 = \frac{2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2}. \tag{7.1}$$

If an internal pilot is also undertaken and the variance, s^2, is re-estimated at the interim. The recalculated sample size, N_{RECALC} will be estimated using the result,

$$N_{RECALC} = \frac{2s^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2}, \tag{7.2}$$

with the procedure described in Figure 7.1. The percentiles of the distribution of σ^2 can be calculated from Equation 7.3 (a restatement of Equation 2.12 presented earlier),

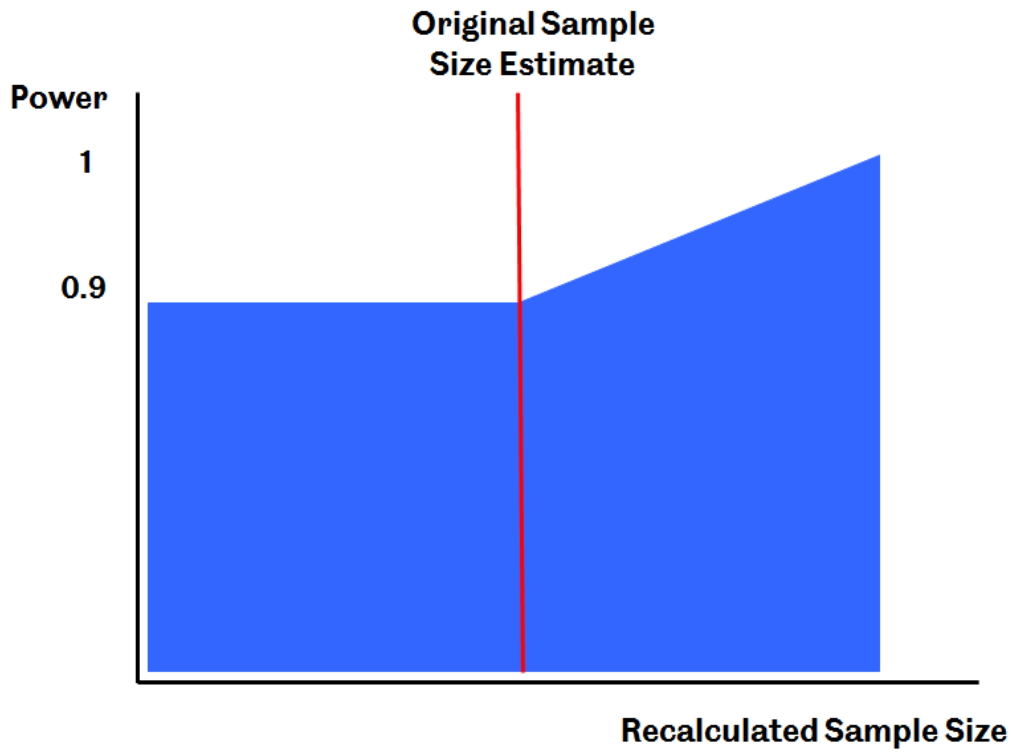
$$\frac{k}{\chi^2_{1-X,k}} s^2, \tag{7.3}$$

(as discussed in Chapter 2) where k is the degrees of freedom for the estimate of σ^2, s^2, and $1 - X$ is the percentile value corresponding to the relevant degrees of freedom of the chi-squared distribution.

The sample size is re-estimated for where the variance, σ^2, used in the initial sample size calculation is assumed known. If the variance is known then all plausible values of s^2 from the distribution of σ^2 for the internal pilot can be estimated using the percentiles of a chi-squared distribution. Equation 7.3 is used to obtain plausible values for s^2 from σ^2. Thus, by investigating the resulting sample size and power level each for the percentiles of the distribution, the average power and sample size which would be achieved through using an internal pilot trial design was calculated.

Figure 7.1 shows that if the recalculated sample size is larger than the original sample size estimate then the power will be larger than the required power (here 0.9) as here we are assuming that the variance is known in the original calculation. If the recalculated sample size is smaller than the original estimate the sample size will not change and the power will be equal to the original required level. This figure demonstrates that the average power will be higher than the original required level as the power will not drop below this assuming that the variance is known in the original calculation, when using an internal pilot trial design

Figure 7.1: Resulting Power if the Original Variance Estimate is Equal to the True Variance



What Figure 7.1 (Table 7.3) is in effect doing is calculating the average power (AP) for a given trial from,

$$AP = P(s_1^2 \geq \sigma^2)\,\Phi\left(\sqrt{\frac{d^2\,(N_1|\,s_1^2,\sigma^2)}{2\sigma^2}} - Z_{1-\alpha/2}\right) + P(s_1^2 < \sigma^2)\,\Phi\left(\sqrt{\frac{d^2\,(N_1|\,\sigma^2)}{2\sigma^2}} - Z_{1-\alpha/2}\right). \quad (7.4)$$

The power for a trial with the sample size $(N_1|\,\sigma^2)$ (the second term on the right hand side) would be equal to 90% (or the specified required power level) as the power here is based on the true variance therefore it can be seen that,

$$AP = P(s_1^2 \geq \sigma^2)\,\Phi\left(\sqrt{\frac{d^2\,(N_1|\,s_1^2,\sigma^2)}{2\sigma^2}} - Z_{1-\alpha/2}\right) + P(s_1^2 < \sigma^2)\,(0.90)\,. \tag{7.5}$$

From Equation 7.2 and Figure 7.1 it can be seen that the internal pilot trial procedure protects against under-powering a trial but can lead to overpowering. Therefore, the Wittes and Brittain (1990) approach can produce powers higher than the nominal level chosen in the sample size calculation, even when the anticipated variance is equal to the true variance, due to its restricted nature in that the sample size can never be adjusted downwards at the interim.

The proportion of times the sample size is increased at the interim can be estimated from $P(s^2 > \sigma^2)$ because the trial is only increased in size if the newly estimated variance s^2 is larger than the original estimate σ^2. As, $\frac{ks^2}{\sigma^2} \sim \chi_k^2$ it can be shown that,

$$\begin{aligned} & P(s^2 > \sigma^2) \\ &= P\left({s^2}/{\sigma^2} > 1\right) \\ &= P\left({ks^2}/{\sigma^2} > k\right) \\ &= P\left(\chi_k^2 > k\right). \end{aligned} \tag{7.6}$$

From this equation is can be seen why the proportion of trials to be increased in sample size tends to 0.5 and increases with the pilot trial sample size. The mean of a chi-squared distribution is the degrees of freedom therefore if the mean and the median are equal the probability above would be equal to 0.5. Therefore, when the number of degrees of freedom is large and the chi-squared distribution tends towards the Normal distribution the proportion of trials increased at the interim tends towards 0.5.

As the pilot trial sample size decreases the skew of the chi-squared distribution increases and the mean will be distorted further towards the right hand tail of the chi-squared distribution. Consequently the probability of $\chi_k^2 > k$ decreases and the proportion of trials which are increased at the interim can be seen to decrease. If the internal pilot trial sample size is fixed no matter the size of the main trial the proportion of the trials to be increased at the interim does not depend on the effect size. For example, where the degrees of freedom for the variance estimate is 18 we have,

$$P(\chi_{18}^2 > 18) = 0.46.$$

However, the degrees of freedom for the estimate of the variance depends on the original sample size calculation (from Equation 7.1) if the internal pilot trial sample size is set proportionate to the size of the main trial. Thus, the probability laid out above depends on the original sample size calculation, where,

$$df = 2\left(\frac{\pi\left(2\sigma^2(Z_{1-\alpha/2} - Z_{1-\beta}\right)^2}{d^2}\right) - 2,$$

and π is the proportion of the main trial to be used as the internal pilot trial sample size. From this equation the proportion of trials that would be increased at the interim can be calculated. If $\sigma^2 = 1$ and $d = 0.05$ then $N_0 = 8406$ and if $\pi = 0.25$ the pilot sample size per arm would be 2,102 therefore we would have 4,202 degrees of freedom for the variance estimate. It therefore follows that the proportion of trials to be increased at the interim will be,

$$P(\chi_{4202}^2 > 4202) = 0.50.$$

This method is implemented by the process outlined in Figure 7.2. The complication in the re-estimation of the sample size is the fact that the procedure is restricted. That is for plausible values of s^2 where the sample size re-estimation is lower than the original calculation we keep the original sample size. The sample size is changed only for plausible values of s^2 where the sample size at the re-estimation is higher than the original calculation, and the sample size is increased to the higher level. This approach is referred to in this thesis as the Wittes and Brittain (1990) approach who outlined the restricted sample size re-estimation procedure, presented in Section 6.3.1.

The nominal power level and the power at the recalculation are set at the 90% level, the Type I error rate was chosen to be a two-sided 5% and the effect size was varied over 0.05, 0.2, 0.5 and 0.8 based with a unit variance. The initial sample size calculation was found based on Equation 7.1, therefore for example $Z_{1-\alpha/2} = 1.96$ and $Z_{1-\beta} = 1.28$. Each percentile of, σ^2 was calculated using Equation 7.3 based on the variance being known and equal to 1 in this situation.

The four options for the size of the internal pilot trial (and hence the degrees of freedom) based on those proposed in the literature and presented in Table 7.1 were investigated. These are 10, $0.25N_0$, $0.5N_0$, and $0.75N_0$. For each percentile value (0.0001 to 0.9999) of the chi-squared distribution the required sample size was recalculated through Equation 7.2 substituting the percentile for, s^2. The restricted procedure was then applied so that the sample size was increased if the new estimate of the sample size was greater than the initial estimate. If the new sample size requirement was lower than the initial calculation, then the initial sample size was used.

Based on the recalculated sample size after applying the restricted approach (N_1) the power of the trial for the true variance was calculated using Equation 7.7 (a restatement of Equation 2.17).

$$1 - \beta = \Phi\left(\sqrt{\frac{d^2 N_1}{2\sigma^2}} - Z_{1-\alpha/2}\right) \quad (7.7)$$

Across the percentiles the results were averaged to find an average sample size and an average power for that design. Additionally, it was recorded when the sample size was increased at the interim, so that it was possible to calculate the overall rate at which the sample size was increased. Altering the variance away from 1 (the value used in these results) would alter the standardised effect size. I have chosen to present the results in terms of the standardised effect size so that the results are more generalizable in practice.

Figure 7.2: Process for Investigating the Effect of an Internal Pilot Trial Design on the Power of the Main Trial

Step 1: Values for Type I (α) and Type II (β) error are selected, the standard deviation value (s) and the effect size (d) is set. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s was set to 1, various values of d were investigated (0.05, 0.2, 0.5 and 0.8). Additionally different sample sizes for the internal pilot trial were investigated namely 20 participants, $0.25N_0$, , $0.5N_0$ and , $0.75N_0$.

Step 2: Calculate , N_0 based on Equation 7.1. Dependent on the rule for size of internal pilot trial being used calculate the sample size for the internal pilot trial.

Step 3: Set i equal to 0.0001, where i represents the percentiles of the chi-squared distribution.

Step 4: For i calculate the percentile of the distribution of σ^2 from Equation 7.3. Where the degrees of freedom are calculated from the sample size of the internal pilot trial.

Step 5: Calculate the required sample size, N_1 based on using this percentile for s^2 through Equation 7.2.

Step 6: Adjust the main trial sample size to N_1 if $N_1 > N_0$ else sample size remains at N_0.

Step 7: Calculate the power of the trial based on the new sample size through Equation 7.4.

Step 8: Record the sample size, power and whether the sample size increased at the interim.

Step 9: If i<0.9999 add 0.0001 to i and go back to Step 4. If i=0.9999 go to Step 10.

Step 10: Find the mean and standard deviation of the sample sizes and powers for all percentiles and the percentage of trials, which were increased in size at the interim.

For example, if the Type I error rate is chosen to be 0.05 or 5% (two-sided), the Type II error rate is set at 0.1 or 10%. If the chosen standardised effect size is 0.5, N_0 would be

set at 170 for a two armed trial. We chose an internal pilot sample size of $0.5N_0$ (hence an internal pilot trial sample size of 86 participants) and set i=0.0001. Following the algorithm for the first 5 loops we get the following values:

i	N_0	N_1	Power	Sample Size was adjusted Upwards at Interim
0.0001	170	170	0.90	No
0.0002	170	170	0.90	No
0.0003	170	170	0.90	No
0.0004	170	170	0.90	No
0.0005	170	170	0.90	No

For the first 5 loops of the algorithm the variance estimated from the distribution is smaller than the true variance used in the original sample size calculation therefore the recalculated sample size at the interim will be smaller than the original calculation. Therefore, the sample size will remain at the original level. For the final 5 loops of the algorithm the variance estimated from the distribution is larger than the true variance used in the original sample size calculation and therefore the sample size at the interim will be adjusted upwards, the results of these loops can be seen below:

i	N_0	N_1	Power	Sample Size was adjusted Upwards at Interim
0.9995	170	268	0.98	Yes
0.9996	170	270	0.98	Yes
0.9997	170	272	0.98	Yes
0.9998	170	276	0.99	Yes
0.9999	170	284	0.99	Yes

After all of the loops of the algorithm have been run we can calculate the average final sample size after the sample size recalculation rules have been applied, the average power of the trials for the true variance and the proportion of trials that were adjusted upwards at the interim sample size recalculation.

The results of this approach are given in Table 7.3, when the size of the internal pilot trial is varied over those presented in Table 7.1. Table 7.2 shows the sample sizes needed for

a conventional sample size with a power level of 90% and a two tailed Type I error rate of 0.05. If there is a sample size re-calculation part way through the trial, for a proportion of the trials the sample size would remain the same (as using a restricted approach) while for some the sample size would increase. Therefore, Table 7.3 gives the impact on the average sample size and power due to the internal pilot trial.

The results show that on average the restricted internal pilot design requires more participants than the fixed sample size design and results in higher average power levels than originally required. This effect is most marked when the sample size of the internal pilot is small. As the internal pilot sample size increases the sample size and power level tend towards those seen in the fixed sample size designs. As the degrees of freedom for the internal pilot variance estimate increase, the results tend to the case where the variance is assumed known throughout (the fixed sample size design).

Table 7.2: Sample Size Requirement for a Two-armed Fixed Sample Size Design with 90% Power

Standardised Effect Size	Sample Size
0.05	16,812
0.20	1,052
0.50	170
0.80	66

Table 7.3: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial Design for a Two Armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.92	0.03	19,025.98	3,673.50	0.46
0.20	0.92	0.03	1,190.23	229.49	0.45
0.50	0.92	0.03	191.74	36.44	0.44
0.80	0.92	0.03	74.95	14.58	0.45
0.25N_0					
0.05	0.90	0.00	16,958.69	216.03	0.50
0.20	0.91	0.01	1,088.46	55.26	0.48
0.50	0.92	0.02	184.16	22.87	0.45
0.80	0.92	0.03	75.47	15.54	0.45
0.5N_0					
0.05	0.90	0.00	16,915.86	152.47	0.50
0.20	0.91	0.01	1,077.75	38.68	0.48
0.50	0.92	0.02	179.90	15.73	0.45
0.80	0.92	0.03	72.83	10.71	0.46
0.75N_0					
0.05	0.90	0.00	16,896.87	124.41	0.50
0.20	0.91	0.01	1,072.98	31.39	0.48
0.50	0.91	0.02	178.01	12.66	0.45
0.80	0.92	0.02	71.65	8.66	0.46

It is of interest in Table 7.3 how the power tends to be greater than the nominal power of 90%. This is because if the variance at the interim is larger than the original estimate the trial sample size will be increased to the recalculated sample size based on the interim estimate of the variance. A consequence is that the trial is now overpowered for the true variance. If, however, the variance at the interim is less than the original variance estimate the trial will carry on with the originally planned sample size. In this situation the power

would remain at 90% power for the true variance. In both cases we are assuming that by the end of the trial the variance estimate will be equal to the true population variance. This idea is also displayed in Figure 7.1.

It is worth noting from Table 7.3 that as the effect size increases the proportion of trials where the sample size is increased at the interim decreases. This is because as the effect size increases the trial sample size decreases and therefore the pilot trial sample size decreases. If the pilot trial sample size used is proportionate to the main trial sample size and the chi-squared distribution used to estimate the plausible values for the interim variance; then as the effect size increases the distribution of the variance becomes more skew this is highlighted below.

7.2.1 Validating the Results through Simulation

To confirm the results simulations were undertaken as follows:

- For the restricted sample size recalculation procedure, the first step is to collect the data in the pilot phase of the trial,
- Using this data, a new estimate of the variance is found and a new estimate of the required sample size is calculated,
- If the recalculated sample size is bigger than the original estimate the sample size for the trial is increased to the new estimate,
- Or if the recalculated sample size is smaller or equal to the original estimate the sample size stays the same as in the original trial sample size calculation.

Figure 7.3 presented below illustrates the method used for the following simulation study looking at the effect of the internal pilot trial design. Note the estimate of the population variance is assumed to be recalculated in a blinded manner as presented in Chapter 6. The result below is used to get an estimate of the variance at the interim (s_1^2),

$$s_1^2 = s_{1,total}^2 - \left(\frac{d}{2}\right)^2$$

where $s_{1,total}^2$ is the estimate of the variance of the trial data ignoring the treatment group and d is the mean difference assumed in the original sample size calculation.

Figure 7.3: Process for Simulating a Trial to Investigate the Effect of the Internal Pilot Trial on the Power of the Main Trial

Step 1: Values for Type I (α) and Type II (β) error are selected, the standard deviation value (s) and the effect size (d) is set. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, s_0 was set to 1, various values of d were investigated (0.05, 0.2, 0.5 and 0.8). Additionally different sample sizes for the internal pilot trial (m) were investigated namely 20, $0.25N_0$, , $0.5N_0$ and , $0.75N_0$.

Step 2: Calculate N_0 based on Equation 7.1.

Step 3: Set i equal to 1, where i represents the number of simulations.

Step 4: For pilot trial sample size, m, simulate the control group from a Normal distribution with variance s_0^2 and a mean of zero in the control group and d in the experimental group.

Step 5: From the simulated data, estimate the blinded variance and recalculate the required sample size N_1, based on this new variance estimate through Equation 7.2.

Step 6: Adjust the main trial sample size to N_1 if $N_1 > N_0$ else sample size remains at N_0.

Step 7: Calculate the power of the trial based on the new sample size through Equation 7.4.

Step 8: Record the sample size, power and whether the sample size increased at the interim.

Step 9: If i is less than the required number of iterations add 1 to i and repeat steps 4 to 9 else continue to Step 10.

Step 10: Find the mean and standard deviation of the sample sizes and powers for all the simulations and the percentage of trials, which were increased in size at the interim.

To choose the required number of iterations for the study first of all the method was run for an effect size of 0.05, nominal required power 0.9, two sided Type I error rate of 0.05, a true variance estimate of 1 and equal allocation of participants between the groups for an internal pilot sample size of $0.25N_0$.

Initially 10,000 trials were simulated and the simulations were run 5 times to check whether the results were stable. The results are shown in Table 7.4.

Table 7.4: Simulations Looking at the Effect of the Internal Pilot Trial Design on the Power and Sample Size of a Trial using 10,000 Iterations

Average Power	Average Sample Size	Standard Deviation	Proportion of Trials Increased at Interim
0.90	16,961.02	216.28	0.50
0.90	16,953.78	211.10	0.49
0.90	16,959.80	218.44	0.49
0.90	16,961.96	222.16	0.50
0.90	16,958.12	215.76	0.50

The results of the simulations seem reasonably stable. To see if the stability is increased the number of iterations is increased to 50,000. The results of these can be seen in Table 7.5.

Table 7.5: Simulations Looking at the Effect of the Internal Pilot Trial Design on the Power and Sample Size of a Trial using 50,000 Iterations

Average Power	Average Sample Size	Standard Deviation	Proportion of Trials Increased at Interim
0.90	16,959.72	216.22	0.50
0.90	16,958.62	217.14	0.49
0.90	16,959.64	216.66	0.50
0.90	16,958.02	216.10	0.50
0.90	16,958.70	216.16	0.50

With 50,000 iterations the simulations seem to be giving more constant results the number of iterations, were increased once again to 100,000 iterations to see if the stability of the results improved any further.

Table 7.6: Simulations Looking at the Effect of the Internal Pilot Trial Design on the Power and Sample Size of a Trial using 100,000 Iterations

Average Power	Average Sample Size	Standard Deviation	Proportion of Trials Increased at Interim
0.90	16,959.00	216.62	0.50
0.90	16,958.80	216.46	0.50
0.90	16,957.04	214.84	0.49
0.90	16,958.30	216.26	0.50
0.90	16,958.04	215.12	0.50

After 100,000 iterations the results for the average power and the proportion of trials to be increased at the interim sample size recalculation were stable and the results for average sample size were stable to the integer value. The number of iterations was increased again to 150,000 to see if the average sample size could be stabilised further.

Table 7.7: Simulations Looking at the Effect of the Internal Pilot Trial Design on the Power and Sample Size of a Trial using 150,000 Iterations

Average Power	Average Sample Size	Standard Deviation	Proportion of Trials Increased at Interim
0.90	16,958.98	215.78	0.50
0.90	16,959.74	216.82	0.50
0.90	16,958.70	216.36	0.50
0.90	16,959.04	217.16	0.50
0.90	16,958.92	216.90	0.50

Increasing the number of iterations to 150,000 did not increase the stability of the average sample size values any further. Thus, it was chosen that the simulations would be based on 100,000 iterations.

The simulation study involves using the true variance of the data as the variance in the original sample size calculation. Recall that the procedure for the simulations is described in Figure 7.3. The parameter values used are α=0.05, β=0.1, r=1, σ^2=1 and d is varied over a range (0.05, 0.2, 0.5 and 0.8). The results, shown in Table 7.8, demonstrate the properties of the internal pilot trial design when the variance used in the sample size calculation is the same as the true variance for the varying size of the internal pilot trial design which can be seen to match closely to those found using the chi-squared distribution presented in Table 7.3.

Table 7.8: Simulations to Show the Properties of Internal Pilot Trial Design Sample Sizes for a Two Armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Proportion of Trials Increased at the Interim
20 participants					
0.05	0.92	0.03	18,971.34	3581.96	0.46
0.20	0.92	0.03	1,187.30	224.96	0.45
0.50	0.92	0.03	192.60	37.62	0.45
0.80	0.93	0.03	76.06	16.32	0.45
0.25N_0					
0.05	0.90	0.00	16,959.70	216.98	0.50
0.20	0.91	0.01	1,088.82	55.86	0.48
0.50	0.92	0.02	184.98	24.14	0.45
0.80	0.93	0.03	76.74	17.46	0.45
0.5N_0					
0.05	0.90	0.00	16,916.78	153.30	0.50
0.20	0.91	0.01	1,078.08	39.08	0.49
0.50	0.92	0.02	180.52	16.64	0.45
0.80	0.92	0.03	73.66	12.10	0.46
0.75N_0					
0.05	0.90	0.00	16,897.12	125.00	0.50
0.20	0.91	0.01	1,073.21	31.72	0.48
0.50	0.91	0.02	178.49	13.39	0.45
0.80	0.92	0.03	72.40	9.76	0.46

7.2.2 The Effect of Using an Adjustment Method at the Sample Size Recalculation

This section aims to investigate the effect in the power and sample size of the adjustment methods used at the interim recalculation. The following results are calculated using the approach set out in Figure 7.2, however after the variance re-estimation the sample size is re-calculated using Equation 4.2 (or Equation 2.15) for the UCL approach and 2.21 for the NCT approach. The parameter values used are α=0.05, β=0.1, σ^2=1, d is varied over a range (0.05, 0.2, 0.5 and 0.8), M_{INT} is the sample size of the pilot trial, which is varied over those presented in Table 7.1 (20 participants, $0.25N_0$, $0.5N_0$ and $0.75N_0$) and where required 1-X= 0.2 (for the 80% UCL approach).

Table 7.9: Properties of the Internal Pilot Trial Design with the NCT Approach at the Sample Size Recalculation Sample Sizes are for a Two-armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.94	0.04	20,768.77	5,035.66	0.62
0.20	0.94	0.03	1,300.24	315.19	0.62
0.50	0.94	0.03	210.25	50.68	0.62
0.80	0.94	0.03	83.17	20.45	0.65
0.25N_0					
0.05	0.90	0.00	16,964.53	219.08	0.51
0.20	0.91	0.01	1,095.06	59.61	0.54
0.50	0.93	0.03	191.70	28.26	0.58
0.80	0.94	0.04	84.74	22.29	0.65
0.5N_0					
0.05	0.90	0.00	16,919.18	154.07	0.51
0.20	0.91	0.01	1,081.47	41.05	0.53
0.50	0.92	0.02	183.96	18.50	0.56
0.80	0.93	0.03	77.63	13.87	0.63
0.75N_0					
0.05	0.90	0.00	16,899.37	125.57	0.51
0.20	0.91	0.01	1,075.76	33.13	0.52
0.50	0.92	0.02	180.99	14.67	0.55
0.80	0.93	0.93	75.14	10.83	0.62

Table 7.10: Properties of the Internal Pilot Trial Design with the 80% UCL Approach at the Sample Size Recalculation Sample Sizes are for a Two-armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.95	0.04	28,173.06	15,793.61	0.65
0.20	0.95	0.04	1,761.88	987.05	0.65
0.50	0.95	0.04	283.04	157.72	0.64
0.80	0.95	0.04	110.80	61.83	0.64
0.25N_0					
0.05	0.91	0.01	17,296.69	547.51	0.66
0.20	0.93	0.03	1,183.37	154.73	0.66
0.50	0.94	0.04	232.14	80.19	0.64
0.80	0.95	0.04	115.01	68.68	0.64
0.5N_0					
0.05	0.91	0.01	17,152.27	382.74	0.66
0.20	0.92	0.02	1,141.86	104.21	0.66
0.50	0.94	0.03	210.02	49.65	0.64
0.80	0.95	0.04	95.80	38.67	0.65
0.75N_0					
0.05	0.90	0.00	17,088.98	310.96	0.66
0.20	0.92	0.02	1,124.21	83.19	0.66
0.50	0.93	0.03	201.33	38.24	0.64
0.80	0.94	0.04	88.90	28.72	0.65

Both the NCT and the UCL approaches make the average powers from the restricted design even higher than when no adjustment method is used given in Table 7.3. A consequence is that more of the trials will be adjusted upwards at the sample size recalculation and the average sample sizes are now higher. The effect is less pronounced

when the pilot trial is larger as the inflation from the adjustment method will be less in these situations.

7.3 Sample Sizes for Internal Pilot Trials – Assuming the Variance is known

It can be seen (in Table 7.3) that as the sample size of the internal pilot increases the effect of the internal pilot trial design on the power of the trial decreases due to the estimate of the variance becoming more precise. Chapter 4 investigated the optimal sample size for an external pilot trial; in this section the required sample size for an internal pilot trial is examined.

Chapter 4 used the overall sample size of the pilot trial and the main trial together to choose a pilot trial sample size that minimises the overall trial sample size. This was based on using an adjustment method for the sample size calculation of the main trial, so that the main trial sample size was dependent on the pilot trial sample size. However, as shown in Section 7.2.2 there is no additional benefit to using an adjustment method at the sample size recalculation in terms of average power.

Figure 7.4 and 7.5 illustrate the sample size calculations for the main trial for different pilot sample sizes (for different effect sizes). Graph A is for an effect size of 0.05. B is for 0.2, C is for 0.5 and D is for an effect size of 0.8. Unlike the external pilot trial situation given in Chapter 4, there is no minimum to the sample size and so as highlighted the bigger the internal pilot the smaller the overall sample size calculation. Additionally, because the internal pilot patients are rolled into the main trial there is no cost (in terms of sample size) to including all the patients in the internal pilot and getting the most accurate prediction of the variance.

For investigators it may not be practical to wait until late into the trial to get the final estimate of the sample size. The requirement for the recalculation to be early in the trial

such that the internal pilot has large enough sample size to have sufficient precision for the re-estimate of the sample size of the main trial needs to be balanced.

Figure 7.4: Average Sample Sizes for the Trial with Varying Internal Pilot Sample Size and Effect Size with 90% Power. A: effect size = 0.05, B: effect size = 0.2, C: effect size = 0.5 and D: effect size = 0.8

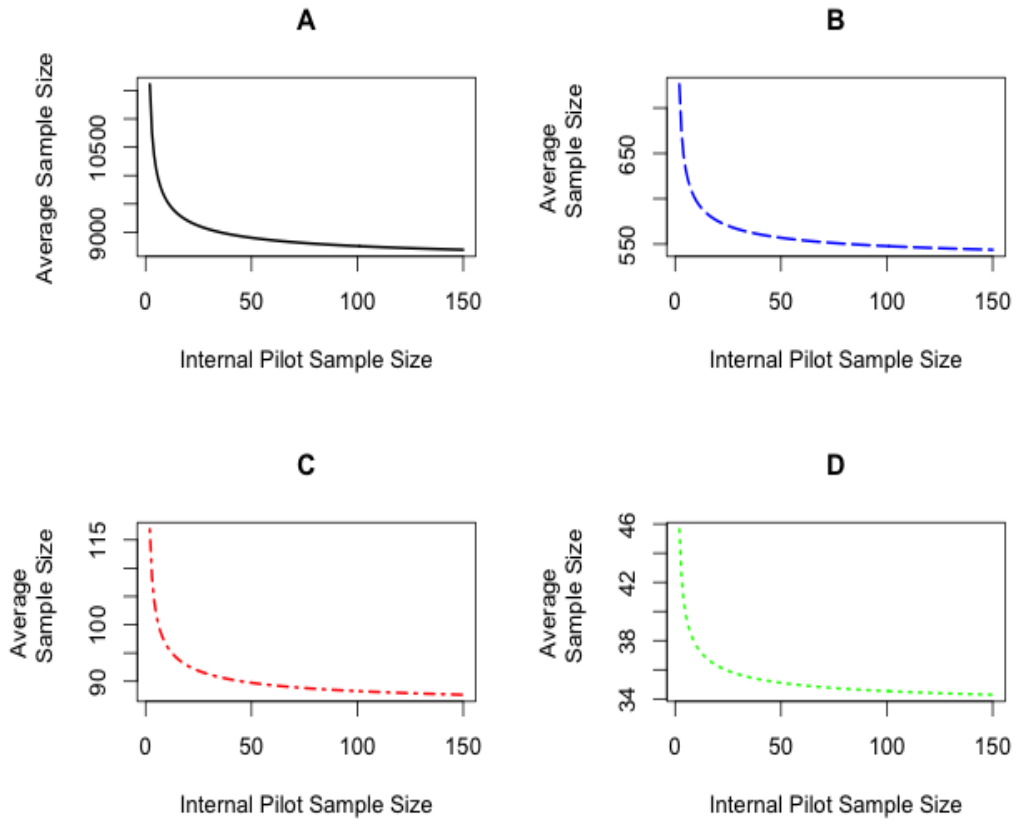


Figure 7.5: Average Sample Sizes for the Trial with Varying Internal Pilot Sample Size and Effect Size with 80% Power. A: effect size = 0.05, B: effect size = 0.2, C: effect size = 0.5 and D: effect size = 0.8

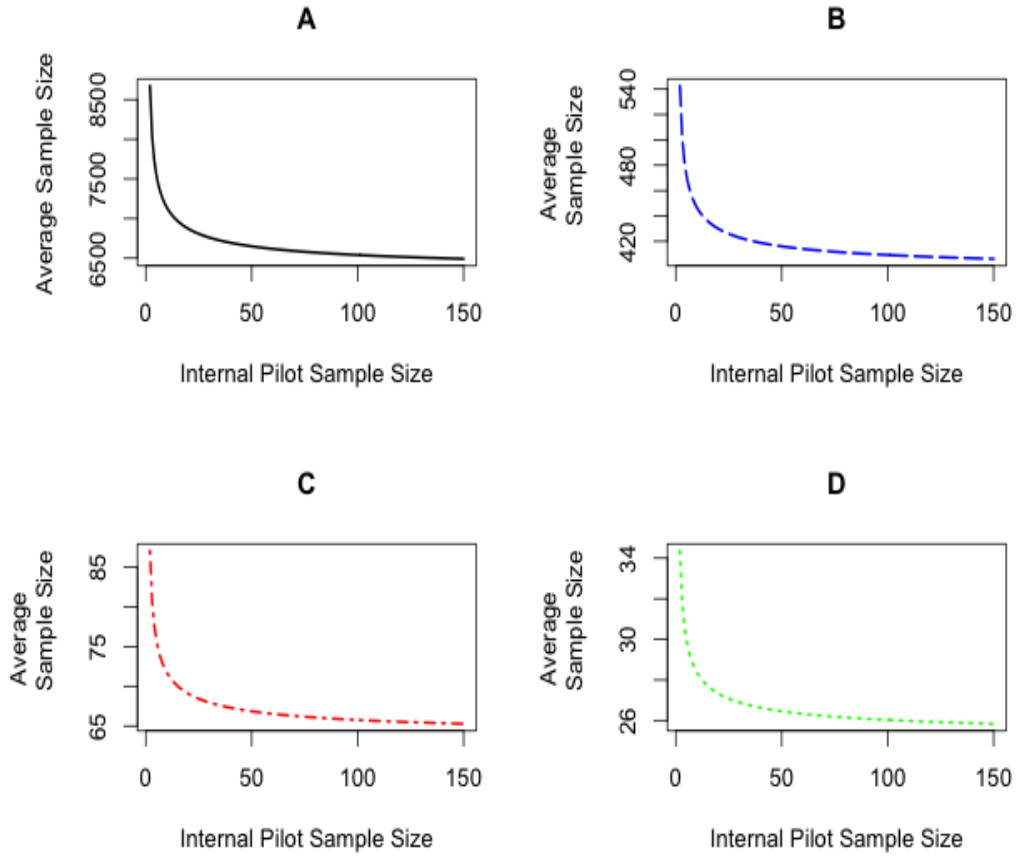


Figure 7.4 and Figure 7.5 can be used to illustrate how the internal pilot trial could be estimated. It can be seen from the figures that the change in the average sample size seems to decrease as the internal pilot trial sample size increases. However, the internal pilot trial sample size has less of an effect on the overall trial sample size, after the internal pilot trial reaches a specific level. This could be taken as when we have a sufficiently precise estimate of the variance.

Suggested estimates for the required size of an internal pilot trial are given in Table 7.11. These were estimated by taking the internal pilot trial sample size to be the point where the absolute change in average sample size of the trial from the compared to previous internal pilot trial sample size drops below two. Two, it should be noted, is arbitrary. As

in Chapter 4 for the external pilot trial sample sizes the sample size per arm has not been allowed to fall below 10 participants. The results in Table 7.11 could be used as the minimum sample size required for an internal pilot trial.

Table 7.11: Sample Size Recommendations for Internal Pilot Trials Assuming the Variance is Known for a Two-armed Trial

Standardised Effect Size	80% Powered Trial	90% Powered Trial
Extra Small ($\delta < 0.1$)	160	190
Small ($0.1 \leq \delta < 0.3$)	30	40
Medium ($0.3 \leq \delta < 0.7$)	20	20
Large ($\delta \geq 0.7$)	20	20

7.4 The Power of a Trial When Using an Internal Pilot Trial – Anticipated Variance is Assumed Known, but is Incorrectly Estimated

The utility of the methods can be investigated for the situation where the assumptions around the variance are wrong. In reality the variance is not known in the initial sample size calculation and so as a result it is possible to under or over-estimate the parameter in the sample size calculation at the start of the trial. This section looks at how the internal pilot trial performs in terms of average power and sample size when the anticipated variance is an over or an under-estimate in the original calculation and compares this to the fixed sample size design. The effects of the adjustment methods are also reinvestigated.

Note the calculations used to estimate the initial sample size are still the methods for the case where the variance is assumed known. Here the situation where the anticipated variance is actually incorrect is explored. Section 7.5 will examine the situation where the population variance is assumed to be known in the initial sample size calculation but is

inaccurately estimated and in reality the estimate of the variance is either larger or smaller than the true value.

7.4.1 Anticipated Variance less than True Variance

If the anticipated variance at the start of a trial were an underestimate the original sample size calculation will be too small hence the trial would be underpowered. Employing an internal pilot trial design is one method to try to protect against this misspecification of the variance in the original calculation.

Table 7.12 shows the effect on the trial power for the true variance and sample size of a fixed trial design if the original variance estimate is 0.75 compared to the true variance of 1. It shows that the power of the trial for the true variance of 1 drops to around 80% as the calculated required sample size falls due to the small estimate of the variance.

Table 7.12: Sample Size Requirement for a Fixed Sample Size Two-armed Design

Standardised Effect Size	Power	Sample Size
0.05	0.80	12,610
0.20	0.80	790
0.50	0.81	128
0.80	0.81	50

In Table 7.12 the sample sizes are smaller than those presented in Table 7.2 which would give 90% power for the true variance. The sample sizes in Table 7.12 are underpowered for the true variance giving only approximately 80% power. Table 7.13 shows the average power of the trials when an internal pilot design is used and the variance is underestimated in the original calculation (anticipated variance=0.75 versus true variance=1). The results show that on average the internal pilot design protects against the under-powering of the trial although the average power can dip slightly below 0.90. When compared to Table 7.3 where the anticipated variance was equal to the true

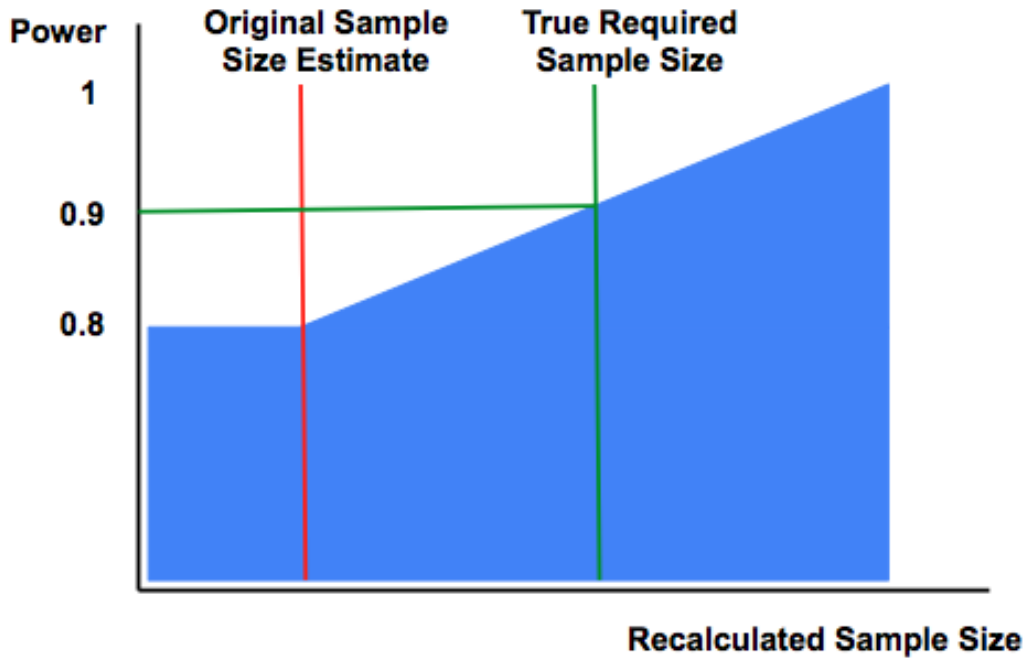
variance the average power and the average sample sizes are lower and just protect against extreme under-powering on average. Moreover, because the original estimate of the variance is an underestimate of the true variance a higher percentage of trials are increased in size at the interim than in Table 7.3.

Table 7.13: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial design with no Adjustment Method for a Two-armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.89	0.07	17,375.33	4,875.37	0.76
0.20	0.89	0.07	1,087.12	304.50	0.76
0.50	0.89	0.06	174.96	48.54	0.75
0.80	0.89	0.06	68.81	19.11	0.75
0.25N_0					
0.05	0.90	0.01	16,812.82	421.18	1.00
0.20	0.90	0.03	1,051.76	105.38	1.00
0.50	0.90	0.06	172.05	39.03	0.82
0.80	0.89	0.07	70.06	22.82	0.69
0.5N_0					
0.05	0.90	0.01	16,812.87	297.81	1.00
0.20	0.90	0.02	1,051.72	74.44	1.00
0.50	0.89	0.05	169.89	28.67	0.92
0.80	0.89	0.06	68.13	16.90	0.79
0.75N_0					
0.05	0.90	0.00	16,812.85	243.16	1.00
0.20	0.90	0.02	1,051.73	60.72	1.00
0.50	0.90	0.04	169.38	23.87	0.96
0.80	0.89	0.05	67.45	14.25	0.85

If at the interim the recalculated variance is less than the original estimate the sample size will remain the same and the trial will be underpowered for the true variance. If at the interim the variance estimate is larger than the anticipated estimate and larger than the true variance, then the sample size will be increased upwards and the trial will be overpowered for the true variance. Additionally, if the variance estimate is larger than the original estimate but smaller than the true variance, the sample size will be increased at the interim but the trial will still be underpowered. This effect is also displayed in Figure 7.6, where the green line represents the required sample size to give 90% power for the true variance.

Figure 7.6: The Effect of Under-Estimating the Variance in the Original Sample Size Calculation



The original sample size estimate is underpowered (for a power requirement of 90%) if the variance is originally underestimated. If the recalculated sample size is to the left of the red line in Figure 7.6 i.e. smaller than the original sample size, the trial will continue to the original sample size so the power of those trials is the power given for the true variance from the original sample size requirement. If the recalculation sample size is to

the right of the line i.e. larger than the original sample size, the sample size will be increased. If the re-estimated variance is still smaller than the true variance the trial will still be underpowered. If the re-estimated variance is larger than the true variance, then the trial will have >90% power.

In Tables 7.14 and 7.15 it can be seen that the adjustment methods give higher powers than when using the internal pilot trial design with no adjustment methods, to protect from under-powering. However, they can lead to over-powering and the UCL approach gives higher powers than the NCT method.

On average the internal pilot trial design protects against under-powering without the adjustment methods, and therefore when the variance is underestimated in the original sample size calculation the adjustment methods offer little more than the internal pilot trial design alone. The restricted method seems to protect against under-powering.

Table 7.14: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial design with the NCT Method used at the Sample Size Recalculation for a Two-armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.91	0.07	19,717.69	6,023.27	0.86
0.20	0.91	0.06	1,235.03	376.60	0.86
0.50	0.92	0.06	200.01	60.40	0.86
0.80	0.92	0.06	79.73	23.87	0.88
0.25N_0					
0.05	0.90	0.01	16,828.58	421.49	1.00
0.20	0.90	0.03	1,067.62	106.87	1.00
0.50	0.91	0.06	187.73	44.71	0.90
0.80	0.93	0.07	85.99	31.34	0.85
0.5N_0					
0.05	0.90	0.01	16,821.70	297.91	1.00
0.20	0.90	0.02	1,060.63	74.93	1.00
0.50	0.91	0.04	178.65	30.57	0.95
0.80	0.92	0.06	76.75	20.08	0.90
0.75N_0					
0.05	0.90	0.00	16,819.42	243.23	1.00
0.20	0.90	0.02	1,058.31	61.01	1.00
0.50	0.91	0.04	175.91	24.80	0.98
0.80	0.91	0.05	73.86	16.01	0.93

Table 7.15: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial design with the 80% UCL Method used at the Sample Size Recalculation for a Two-armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.92	0.08	26,972.47	16,731.93	0.78
0.20	0.92	0.08	1,686.92	1,045.61	0.78
0.50	0.92	0.08	270.91	167.16	0.78
0.80	0.93	0.08	106.29	65.38	0.78
0.25N_0					
0.05	0.91	0.01	17,188.74	861.22	1.00
0.20	0.91	0.05	1,164.23	229.44	0.96
0.50	0.92	0.07	236.76	112.84	0.81
0.80	0.93	0.08	122.92	93.55	0.75
0.5N_0					
0.05	0.90	0.01	17,074.58	604.90	1.00
0.20	0.91	0.04	1,125.15	158.74	0.99
0.50	0.92	0.06	208.54	70.00	0.87
0.80	0.92	0.07	98.02	52.08	0.79
0.75N_0					
0.05	0.90	0.01	17,025.08	492.46	1.00
0.20	0.91	0.03	1,109.83	128.16	1.00
0.50	0.92	0.06	198.52	54.81	0.90
0.80	0.92	0.07	89.56	38.91	0.82

7.4.2 Anticipated Variance larger than True Variance

If the variance estimate at the start of a trial is an overestimate the original sample size calculation will be too large hence the trial would be overpowered. Table 7.16 shows the effect on the power for the true variance and sample size of a fixed trial design if the original variance estimate is 1.5 compared to the variance of 1. It shows that the power of the trial for the true variance of 1 rises to around 98% as the calculated required sample size increases due to the large estimate of the variance.

Table 7.16: Sample Size Requirement for a Two-armed Trial Fixed Sample Size Design

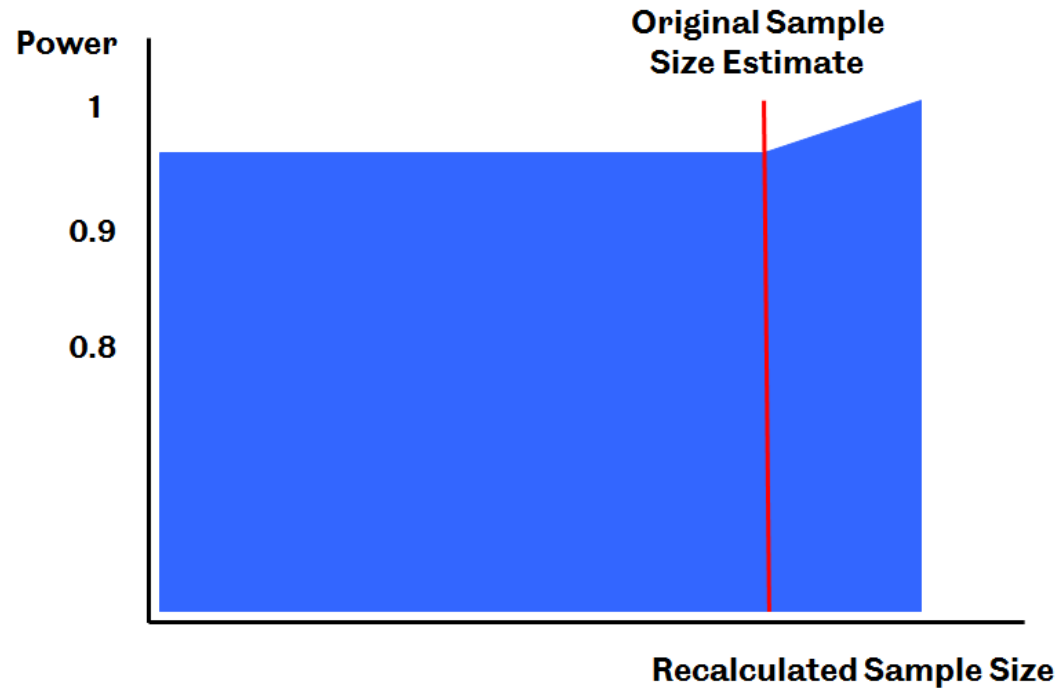
Standardised Effect Size	Power	Sample Size
0.05	0.98	25,218
0.20	0.98	1,578
0.50	0.98	254
0.80	0.98	100

In Table 7.16 the sample sizes are larger than those presented in Table 7.2 which would give 90% power for the true variance. The sample sizes in Table 7.16 are overpowered for the true variance giving approximately 98% power. Table 7.17 shows the average power of the trials when an internal pilot design is used and the variance is overestimated in the original calculation. The results show that the restricted internal pilot trial design leads to overpowering with an average power of 98%, the same as in Table 7.16.

With the restricted design if the variance at the sample size recalculation is less than the original estimate and less than the true variance, the trial will continue with the original sample size calculation and it will be overpowered for the true variance. If the variance at the sample size recalculation is less than the original estimate and more than the true variance the trial will continue and it will be overpowered for the true variance. If the variance at the sample size recalculation is more than the original estimate, then the

sample size will be increased and the trial will be further overpowered for the true variance. This effect is also demonstrated in Figure 7.7.

Figure 7.7: The Effect of Over-Estimating the Variance in the Original Sample Size Calculation



The variance in the original calculation is too high therefore because of the restricted nature of this internal pilot trial procedure the trials will all be overpowered for the true variance. The trials where the re-calculated variance is less than the original estimate will carry on to the original sample size level and hence be overpowered. If the recalculated sample size is to the right of the line i.e. larger than the original estimate, then the sample size will be adjusted upwards and the trial will be over-powered.

Table 7.17: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial Design with no Adjustment Method Applied for a Two-armed Trial

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.98	0.00	25,490.56	1,271.24	0.08
0.20	0.98	0.00	1,594.97	79.27	0.08
0.50	0.98	0.00	256.66	12.54	0.08
0.80	0.98	0.00	101.03	4.87	0.07
0.25N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	254.06	0.99	0.01
0.80	0.98	0.00	100.56	3.21	0.05
0.5N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	254.00	0.00	0.00
0.80	0.98	0.00	100.07	0.82	0.02
0.75N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	254.00	0.00	0.00
0.80	0.98	0.00	100.01	0.20	0.00

Table 7.18: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial design with the NCT Method used at the Sample Size Recalculation for Two-armed Trials

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.98	0.00	26,013.50	2,395.53	0.18
0.20	0.98	0.00	1,627.88	149.98	0.17
0.50	0.98	0.00	262.16	24.26	0.18
0.80	0.98	0.00	103.38	9.77	0.18
0.25N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	254.18	2.00	0.01
0.80	0.98	0.00	101.78	6.31	0.12
0.5N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	254.00	0.00	0.00
0.80	0.98	0.00	100.23	1.66	0.03
0.75N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	254.00	0.00	0.00
0.80	0.98	0.00	100.04	0.49	0.01

Table 7.19: Average Power, Sample Size and Percentage of Increases in Sample Size at Interim when using the Restricted Internal Pilot Trial design with the 80% UCL Method used at the Sample Size Recalculation for a Two-armed Trials

Effect Size	Average Power	Standard Deviation	Average Sample Size	Standard Deviation	Percentage of Trials which Increased at Interim
20 participants					
0.05	0.98	0.01	32,161.98	13,382.74	0.41
0.20	0.98	0.01	2,011.64	836.15	0.41
0.50	0.99	0.01	323.10	133.59	0.41
0.80	0.99	0.01	126.90	52.09	0.41
0.25N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.10	5.16	0.01
0.50	0.98	0.01	267.11	34.74	0.22
0.80	0.98	0.01	119.06	38.48	0.37
0.5N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	257.30	13.16	0.10
0.80	0.98	0.00	107.56	18.23	0.26
0.75N_0					
0.05	0.98	0.00	25,218.00	0.00	0.00
0.20	0.98	0.00	1,578.00	0.00	0.00
0.50	0.98	0.00	255.10	6.32	0.05
0.80	0.98	0.00	103.77	10.89	0.19

In Tables 7.18 and 7.19 it can be seen that the adjustment methods offer no protection against overpowering the trials and require higher average sample sizes than the unadjusted approach.

After reviewing all of the situations that may arise when estimating the variance at the start of a trial and then carrying out a sample size recalculation part way through the trial there does not seem to be any benefit in using an adjustment method at the sample size recalculation as part of an internal pilot trial design. The internal pilot trial design itself protects from under-powering when the original variance estimate is too small compared to the true population variance.

7.5 The Sample Size and Power of a Trial When Using an Internal Pilot Trial – Allowing for Unknown Variance with Pilot Sample Size Fixed

In contradiction to the methods laid out in Section 7.2, where we assume the variance is known the initial sample size calculation will most likely be based on some estimate of the variance say, s_0^2, which may come from an external pilot trial for example. Substituting in s_0^2 for σ^2 into the original sample size formula gives,

$$N_0 = \frac{2s_0^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2}. \tag{7.8}$$

When the variance is re-estimated at the interim the sample size recalculation will be based on,

$$N_0 = \frac{2s_1^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2}, \tag{7.9}$$

where s_1^2 is the re-estimated variance from the internal pilot trial. It is known that,

$$\frac{ks^2}{\sigma^2} \sim \chi_k^2. \tag{7.10}$$

Therefore, $k_0 s_0^2/\sigma^2 \sim \chi^2_{k_0}$ where k_0 is the degrees of freedom for s_0^2 the initial variance estimate and, $k_1 s_1^2/\sigma^2 \sim \chi^2_{k_1}$ where k_1 is the degrees of freedom for s_1^2 the variance estimate from the internal pilot trial.

The ratio of two chi-squared distributions over their respective degrees of freedom follows a F-distribution (Hiorns, 1971) (as discussed in Section 2.3.4) i.e.,

$$\begin{aligned} &\frac{k_0 s_0^2/\sigma^2}{k_0} \Big/ \frac{k_1 s_1^2/\sigma^2}{k_1} \sim F(k_0, k_1) \\ &\Rightarrow \frac{s_0^2}{\sigma^2} \Big/ \frac{s_1^2}{\sigma^2} \sim F(k_0, k_1) \\ &\Rightarrow \frac{s_0^2}{s_1^2} \sim F(k_0, k_1) \end{aligned} \tag{7.11}$$

The average power of the trial would now be calculated from

$$\begin{aligned} AP = P(s_1^2 \geq s_0^2)\, \Phi\left(\sqrt{\frac{d^2\,(N_1 |\, s_1^2, s_0^2)}{2\sigma^2}} - z_{1-\alpha/2}\right) \\ + P(s_1^2 < s_0^2)\, \Phi\left(\sqrt{\frac{d^2\,(N_1 |\, s_0^2)}{2\sigma^2}} - z_{1-\alpha/2}\right), \end{aligned} \tag{7.12}$$

where,

$$\begin{aligned} &P(s_1^2 \geq s_0^2) \\ &= P\left(\frac{s_0^2}{s_1^2} \leq 1\right) \\ &= P(F(k_0, k_1) \leq 1), \end{aligned}$$

and,

$$\begin{aligned} &P(s_1^2 < s_0^2) \\ &= P\left(\frac{s_0^2}{s_1^2} > 1\right) \end{aligned}$$

$$= P(F(k_0, k_1) > 1)$$
$$= 1 - P(F(k_0, k_1) \leq 1).$$

The average sample size of the trial would now be calculated from:

$$ASS = P(s_1^2 \geq s_0^2)\,(N_1 |\, s_1^2, s_0^2) + P(s_1^2 < s_0^2)\,(N_1 |\, s_0^2) \tag{7.13}$$

$$ASS = P(F(k_0, k_1) \leq 1)\,(N_1 |\, s_1^2, s_0^2) + 1 - P(F(k_0, k_1) \leq 1)\,(N_1 |\, s_0^2)$$

The next set of calculations presented in Table 7.20 look at the effect of allowing for the variance being unknown on the sample size of the trial. These results were obtained by the process set out in Figure 7.8; the percentiles were reduced to three decimal places to reduce the run time of the computer program compared to the procedure outlined in Figure 7.1, which was carried out with the percentiles to four decimal places.

This process is an extension of the process carried out earlier in the Chapter and presented in Figure 7.1, which investigated the effect of an internal pilot trial design on the power of the main trial. However, previously it was assumed that the variance at the start of the trial was known and only the variance at the interim was an estimate.

The extension presented here also allows for the variance to be an estimate at the start of the trial in the original sample size calculation as well as at the interim sample size recalculation. Therefore, this algorithm has a double loop, one for evaluating over the distribution for the prior estimate of the variance (from perhaps an external pilot trial) and the other over the distribution for the estimate of the variance form the internal pilot trial.

Figure 7.8: Process for Investigating the Effect of an Estimated Variance in the Initial Sample Size Calculation

Step 1: Values for Type I (α) and Type II (β) error are selected, the true variance (σ^2) and the effect size (d) is set. Here $\alpha = 0.05$ (two-sided), $\beta = 0.1$, σ^2 was set to 1, various values of d were investigated (0.05, 0.2, 0.5 and 0.8). Additionally, different sample sizes for the external and internal pilot trial can be investigated.

Step 2: Set j equal to 0.001, this represents the percentile of the distribution of the prior estimate of the variance.

Step 3: For j calculate the percentile of the distribution of σ^2 through Equation 7.3.

Step 4: Calculate the required sample size, N_0 through Equation 7.9 based on the percentile calculated in Step 3.

Step 5: Using this N_0 as the N_0 in step 1 of the process presented in Figure 7.1. Carry out the procedure to get an average trial power and average sample size over the distribution of the variance in the internal pilot trial, for this percentile of the distribution for the prior estimate of the variance.

Step 6: This should be repeated for each percentile of the distribution of the prior estimate of the variance. Therefore, is j<0.999 and 0.001 to j and go back to Step 3. Otherwise go to Step 7.

Step 7: Find the mean and standard deviation of the average sample sizes and powers for all the percentiles.

For example, if the Type I error rate is chosen to be 0.05 or 5% (two-sided), the Type II error rate is set at 0.1 or 10%. If the chosen standardised effect size is 0.5 and we chose an external pilot trial of 100 and internal pilot sample size of 20 participants. Using a true variance value of 1. Following the algorithm for the final percentile of the distribution of the variance based on the prior estimate, the first 5 percentiles for the variance based on the internal pilot trial would yield the following values:

Percentile of distribution of variance from the Internal Pilot Trial	Average Trial Sample Size	Average Trial Power
0.001	84.56	0.90
0.002	84.56	0.90
0.003	84.56	0.90
0.004	84.56	0.90
0.005	84.56	0.90

All the percentiles of this distribution are calculated, and this is repeated for al percentiles of the distribution of the prior variance. The final 5 percentiles give the following results:

Percentile of distribution of variance from the Internal Pilot Trial	Average Trial Sample Size	Average Trial Power
0.995	119.00	0.97
0.996	120.00	0.97
0.997	121.00	0.97
0.998	123.00	0.98
0.999	127.00	0.98

Running the whole of the procedure would give an average final trial sample size of 96.53 and an average power of 92%.

The internal pilot trial sample size has been set at the recommended size as given in Section 7.3, Table 7.11. The results show that for small sample sizes for the external pilot trial a higher sample size on average is required to maintain the same trial power as in the variance known case. The results tend to the variance known case as the sample size of the external pilot trial increases.

Sample sizes calculations depend on multiple parameters many of which have been generalised here in order to inform investigators how the methods they select affect the power and sample size of the main trial. The methods described can be generalised to a

sample size estimate for a given external pilot variance estimate and a fixed internal pilot sample size.

The internal pilot trial sample sizes have been fixed in these investigations to allow the degrees of freedom for both s_0^2 and s_1^2 to be known and therefore the properties of the χ^2 distribution to be used to study the effect on the expected power and sample size of the main trial. However, using a proportional rule to choose the pilot trial sample size would mean that s_1^2 is dependent on s_0^2. It is unknown if or how this would affect the distribution of the variances from the pilot trials.

Table 7.20: The Effect on the Sample Size and Power of Allowing for the Variance being an Estimate

External Pilot Trial Sample Size	Internal Pilot Trial Sample Size	Average Final Trial Sample Size	Standard Deviation	Average Power	Standard Deviation
d=0.05					
Variance Known	95	8,750.39	522.22	0.91	0.01
10	95	9,559.11	1,796.22	0.92	0.03
25	95	9,166.55	1,038.32	0.92	0.02
50	95	8,993.06	698.04	0.91	0.02
100	95	8,886.28	472.79	0.91	0.01
1000	95	8,766.33	137.98	0.91	0.00
d=0.2					
Variance Known	20	573.59	75.70	0.92	0.02
10	20	609.24	109.42	0.92	0.03
25	20	589.47	62.52	0.92	0.03
50	20	581.92	42.08	0.92	0.02
100	20	577.85	28.84	0.92	0.02
1000	20	573.97	8.81	0.92	0.00
d=0.5					
Variance Known	10	95.82	18.04	0.92	0.03
10	10	100.04	17.27	0.92	0.04
25	10	97.45	9.92	0.92	0.03
50	10	96.53	6.75	0.92	0.02
100	10	96.06	4.66	0.92	0.01
1000	10	95.62	1.45	0.92	0.00
d=0.8					
Variance Known	10	37.46	7.22	0.92	0.03
10	10	39.38	6.75	0.92	0.04
25	10	38.37	3.88	0.92	0.03
50	10	38.01	2.64	0.92	0.02
100	10	37.83	1.83	0.92	0.01
1000	10	37.66	0.59	0.93	0.00

Table 7.21 to 7.24 shows the average sample sizes and powers for trials with an internal pilot trial design with standardised effect sizes of 0.05, 0.2, 0.5 and 0.8. When the sample sizes for the two internal and external pilot trials are varied over a range. The same process as presented above is used however this time the internal pilot sample size is also varied over a range. As the two variances are assumed to be independent random samples from the population the tables are symmetric along the diagonal.

Table 7.21: Average Sample Size and Power for Trials with an Internal Pilot Trial and Standardised Effect Size of 0.05 Allowing for the Variance to be an Estimate in the Original Calculation for a Two-armed Trial

		IPT Sample Size					
		20	**40**	**60**	**100**	**200**	**1000**
EPT Sample Size	**20**	19,909.32 0.92	19,480.72 0.92	19,331.00 0.92	19,207.86 0.92	19,113.18 0.92	19,035.80 0.92
	40	19,480.73 0.92	18,961.45 0.92	18,769.92 0.92	18,606.52 0.92	18,476.26 0.92	18,366.10 0.92
	60	19,331.01 0.92	18,769.92 0.92	18,556.54 0.92	18,370.05 0.92	18,217.31 0.92	18,084.31 0.92
	100	19,207.87 0.92	18,606.52 0.92	18,370.05 0.92	18,157.10 0.92	17,975.98 0.91	17,810.64 0.91
	200	19,113.17 0.92	18,476.26 0.92	18,217.31 0.92	17,975.98 0.91	17,760.08 0.91	17,546.77 0.91
	1000	19,035.79 0.92	18,366.10 0.92	18,084.31 0.92	17,810.64 0.91	17,546.77 0.91	17,235.38 0.91

Table 7.22: Average Sample Size and Power for Trials with an Internal Pilot Trial and Standardised Effect Size of 0.2 Allowing for the Variance to be an Estimate in the Original Calculation for a Two-armed Trial

		IPT Sample Size					
		20	**40**	**60**	**100**	**200**	**1000**
	20	1,243.38	1,216.90	1,207.74	1,200.30	1,194.64	1,190.08
EPT		0.92	0.92	0.92	0.92	0.92	0.92
Sample							
Size	**40**	1,216.90	1,184.56	1,172.72	1,162.70	1,154.78	1,148.16
		0.92	0.92	0.92	0.92	0.92	0.92
	60	1,207.74	1,172.72	1,159.48	1,147.98	1,138.62	1,130.58
		0.92	0.92	0.92	0.92	0.92	0.92
	100	1,200.30	1,162.70	1,147.98	1,134.78	1,123.62	1,113.56
		0.92	0.92	0.92	0.92	0.91	0.91
	200	1,194.64	1,154.78	1,138.62	1,123.62	1,110.24	1,097.14
		0.92	0.92	0.92	0.91	0.91	0.91
	1000	1,190.08	1,148.16	1,130.58	1,113.56	1,097.14	1,077.82
		0.92	0.92	0.92	0.91	0.91	0.91

Table 7.23: Average Sample Size and Power for Trials with an Internal Pilot Trial and Standardised Effect Size of 0.5 Allowing for the Variance to be an Estimate in the Original Calculation for a Two-armed Trial

		IPT Sample Size					
		20	**40**	**60**	**100**	**200**	**1000**
	20	200.08	195.80	194.29	193.07	192.12	191.35
EPT		0.92	0.92	0.92	0.92	0.92	0.92
Sample							
Size	**40**	195.80	190.61	188.68	187.06	185.75	184.66
		0.92	0.92	0.92	0.92	0.92	0.92
	60	194.29	188.68	186.54	184.69	183.15	181.63
		0.92	0.92	0.92	0.92	0.92	0.92
	100	193.07	187.06	184.69	182.57	180.75	179.10
		0.92	0.92	0.92	0.92	0.92	0.91
	200	192.12	185.75	183.15	180.75	178.58	176.45
		0.92	0.92	0.92	0.92	0.91	0.91
	1000	191.35	184.66	181.63	179.10	176.45	173.35
		0.92	0.92	0.92	0.91	0.91	0.91

Table 7.24: Average Sample Size and Power for Trials with an Internal Pilot Trial and Standardised Effect Size of 0.8 Allowing for the Variance to be an Estimate in the Original Calculation for a Two-armed Trial

		IPT Sample Size					
		20	**40**	**60**	**100**	**200**	**1000**
	20	78.77	77.09	76.51	76.03	75.66	75.35
EPT		0.92	0.92	0.92	0.92	0.92	0.93
Sample							
Size	**40**	77.09	75.06	74.32	73.68	73.17	72.74
		0.92	0.92	0.92	0.92	0.92	0.92
	60	76.51	74.32	73.49	72.76	72.16	71.64
		0.92	0.92	0.92	0.92	0.92	0.92
	100	76.03	73.68	72.76	71.92	71.21	70.64
		0.92	0.92	0.92	0.92	0.92	0.92
	200	75.66	73.17	72.16	71.21	70.36	69.53
		0.92	0.92	0.92	0.92	0.92	0.91
	1000	75.35	72.74	71.64	70.64	69.53	68.32
		0.93	0.92	0.92	0.92	0.91	0.91

7.6 Altering the Required Power Levels

As shown in Table 7.20 the internal pilot trial design leads to a slight inflation in the average power of the trial. Changing the nominal power in the sample size calculation is analogous to over or underestimating the variance in the original calculation, in that it would alter the sample size to be required and therefore lead to the situations described in Figures 7.6 and 7.7. Reducing the required power would reduce the initial sample size requirement and result in a situation like in Figure 7.6.

Figure 7.6 shows that if the original sample size is an underestimate then some of the trials will be underpowered. Where the re-estimated variance is higher than the true variance some of the trials will be overpowered. Therefore, the underpowered trials offset some of the overpowered trials to bring the average power of the trials down. It should be possible for a given study to choose the nominal power level such that the average power is the required level.

7.6.1 Assuming the Variance is Known

If the variance at the start of the trial is assumed to be known, then the average power can be calculated from,

$$AP = P(s_1^2 \geq \sigma^2)\, E(1-\beta|s_1^2, (1-\beta)_1) + P(s_1^2 < \sigma^2)\, E(1-\beta|\sigma^2, (1-\beta)_0), \quad (7.14)$$

and the average sample size can be calculated from,

$$ASS = P(s_1^2 \geq \sigma^2)\, E(N_1|s_1^2, (1-\beta)_1) + P(s_1^2 < \sigma^2)\, E(N_1|\sigma^2, (1-\beta)_0), \quad (7.15)$$

where $(1-\beta)_1$ is the power level set at the sample size recalculation and $(1-\beta)_0$ is the power level set at the initial sample size calculation and therefore the results in Table 7.20 for the variance known case can be derived.

If the standardised effect size is 0.2 and the internal pilot trial sample size is set at 40 participants across two treatment arms. Then using the chi-squared distribution and the degrees of freedom it can be shown that (based on Equation 7.14),

$$AP = P\left(\chi_k^2 \geq k\right) E(1-\beta|s_1^2, \sigma^2, (1-\beta)_1) + P\left(\chi_k^2 < k\right) E(1-\beta|\sigma^2, (1-\beta)_0)$$

$$AP = P\left(\chi_k^2 \geq k\right) \int_{P\left(s_1^2 < \sigma^2\right)}^{1} (1-\beta|s_1^2, \sigma^2, (1-\beta)_1)\, ds_1^2 + P\left(\chi_k^2 < k\right) \int_0^{P\left(s_1^2 < \sigma^2\right)} (1-\beta|\sigma^2, (1-\beta)_0)\, d\sigma^2$$

$$AP = 0.469 \int_{0.531}^{1} (1-\beta|s_1^2, \sigma^2, (1-\beta)_1)\, ds_1^2 + 0.531 \int_0^{0.531} (1-\beta|\sigma^2, (1-\beta)_0)\, d\sigma^2$$

$$AP = 0.469\ (0.938) + 0.531\ (0.9)$$
$$AP = 0.440 + 0.478$$
$$AP = 0.918 \sim 0.92.$$

This can be seen to match the result displayed in Table 7.20. The average sample size can also be calculated from,

$$ASS = P\left(\chi_k^2 \geq k\right) E(N_1|s_1^2, \sigma^2, (1-\beta)_1) + P\left(\chi_k^2 < k\right) E(N_1|\sigma^2, (1-\beta)_0)$$

$$ASS = 0.469 \int_{0.531}^{1} (N_1|s_1^2, \sigma^2, (1-\beta)_1 \, ds_1^2 + 0.531 \int_{0}^{0.531} (N_1|\sigma^2, (1-\beta)_0) \, d\sigma^2$$

$$ASS = 0.469\,(627.36) + 0.531\,(526)$$
$$ASS = 294.23 + 279.31$$
$$ASS = 573.54.$$

This result approximately (to one decimal place) matches the result in Table 7.20 the mismatch is due to rounding and the finite limit of percentile that were investigated. It was shown that the internal pilot trial procedure leads to higher power on average than originally required. From the equations above it can be seen that if the required average power is specified it is possible to alter the nominal or the recalculation power to make the average power equal the required level. For example, if the standardised effect size is equal to 0.2,

$$AP = 0.469 \int_{0.531}^{1} (1-\beta|s_1^2, \sigma^2, (1-\beta)_1) \, ds_1^2 + 0.531 \int_{0}^{0.531} (1-\beta|\sigma^2, (1-\beta)_0) \, d\sigma^2$$

$$AP = 0.469\,(0.938) + 0.531\,((1-\beta)_0)$$

Therefore, if the required average power is specified the formula can be re-arranged to find the required nominal power to end with this average power overall. For example, if the required average power is 0.90 then,

$$0.90 = 0.469\,(0.938) + 0.531\,((1-\beta)_0)$$
$$0.90 = 0.440 + 0.531\,((1-\beta)_0)$$
$$\frac{0.90 - 0.440}{0.531} = (1-\beta)_0$$
$$(1-\beta)_0 = 0.866.$$

Consequently, if the power in the original sample size calculation had been set at 86.6% the average power of the trial would have been 90%. Alternatively, the recalculation power could be manipulated to reduce the power of the trials, which are increased in size at the interim recalculation.

7.6.2 Allowing the Variance to be Unknown

If the variance at the start of the trial is assumed to be unknown then the average power would be calculated from,

$$AP = P(s_1^2 \geq s_0^2)\, E(1-\beta | s_0^2, s_1^2, (1-\beta)_1) + P(s_1^2 < s_0^2)\, E(1-\beta | s_0^2, (1-\beta)_0), \quad (7.16)$$

and the average sample size would be equal to,

$$ASS = P(s_1^2 \geq s_0^2)\, E(N_1 | s_0^2, s_1^2, (1-\beta)_1) + P(s_1^2 < s_0^2)\, E(N_1 | s_0^2, (1-\beta)_0). \quad (7.17)$$

These equations become Equations 7.18 and 7.19 when the probability statements about s_0^2 and s_1^2 (from Section 7.5) are replaced with statements of probability about an F-distribution,

$$AP = P(F(k_0, k_1) \leq 1)\, E(1-\beta | s_0^2, s_1^2, (1-\beta)_1) + \big(1 - P(F(k_0, k_1) \leq 1)\big)\, E(1-\beta | s_0^2, (1-\beta)_0), \quad (7.18)$$

and,

$$ASS = P(F(k_0, k_1) \leq 1)\, E(N_1 | s_0^2, s_1^2, (1-\beta)_1) + \left(1 - P(F(k_0, k_1) \leq 1)\right) E(N_1 | s_0^2, (1-\beta)_0), \tag{7.19}$$

These expectations are derived using the formula presented below with bounds based on the probabilities from the F-distribution,

$$AP = P(F(k_0, k_1) \leq 1) \int_0^1 \int_{P\left(s_1^2 < s_0^2\right)}^1 (1 - \beta | s_0^2, s_1^2, (1-\beta)_1)\ ds_1^2 ds_0^2\ + \left(1 - P(F(k_0, k_1) \leq 1)\right) \int_0^{P\left(s_1^2 < s_0^2\right)} (1 - \beta | s_0^2, (1-\beta)_0)\ ds_0^2, \tag{7.20}$$

and,

$$ASS = P(F(k_0, k_1) \leq 1) \int_0^1 \int_{P\left(s_1^2 < s_0^2\right)}^1 (N_1 | s_0^2, s_1^2, (1-\beta)_1)\ ds_1^2 ds_0^2\ + \left(1 - P(F(k_0, k_1) \leq 1)\right) \int_0^{P\left(s_1^2 < s_0^2\right)} (N_1 | s_0^2, (1-\beta)_0)\ ds_0^2, \tag{7.21}$$

from these formula the rest of the results in Table 7.20 could be calculated. However, in the previous section the integrals were bounded by the probability $P(s_1^2 < \sigma^2)$, because σ^2, the true variance is fixed it is possible to calculate this probability which depends on a chi-squared distribution and its degrees of freedom. For the variance unknown case the integral will be bounded by the probability $P(s_1^2 < s_0^2)$. Because the expectation is being taken over both s_0^2 and s_1^2 the probability will change with each percentile of s_0^2 (i.e. for every percentile of s_0^2 there will be a different value for $P(s_1^2 < s_0^2)$) therefore, the calculation is iterative and cannot be solved using the same process as in the variance known case.

7.7 Summary

The internal pilot trial procedure gives trial powers higher than the required level even when the variance used in the original calculation is equal to the true variance. When the variance in the original sample size calculation is too low the internal pilot trial design on average protects against the under-powering, which would normally result. However, if the variance in the original sample size calculation is too high the restricted design leads to very high average sample sizes and powers. The adjustment methods previously discussed in Chapter 4 offer no help to deal with this overpowering. The internal pilot trial also deals with the problem of under-powering inherently therefore there is no extra benefit to the average power of the trial by using an adjustment method at the interim sample size recalculation.

Sample size recommendations were given for when the variance in the original calculation is assumed to be known. These were based on selecting a pilot trial sample size where the change in overall trial sample size falls below 2 participants when the pilot trial sample size is increased any further. This method of selecting the pilot trial sample size was necessary due to the fact that unlike in previous chapters regarding external pilot trials the overall trial sample size does not have a minimum, there is no penalty in terms of numbers of participants required in including all patients in the internal pilot to get the best estimate of the variance.

The chapter also investigated how allowing for the fact that the variance is an estimate, perhaps from an external pilot trial, affects the power and required sample size of the trial. Having an external pilot as well as an internal pilot does increase the average sample size. The sample sizes converge to the case of known variance in the initial sample size calculation as the external pilot sample size increases.

Throughout this chapter it has been observed that the internal pilot trial design leads to higher than required powers. Finally, the chapter investigated how the nominal or

recalculation power might be altered to bring the average power to equal the required level; and how this new power level might be calculated if the variance is assumed to be known in the initial sample size calculation.

Chapter 8

Discussion

8.1 Introduction

This thesis investigated the required sample size for pilot trials for the situations where we have both external and internal pilot designs. It looked to take into account the imprecision involved in estimating the variance from a pilot trial to minimise the overall sample size of the pilot and the main trial together while maintaining the power and Type I error rate.

The sample size is an important consideration when planning a clinical trial. When the outcome is a continuous variable, part of the calculation of the sample size requires an accurate estimate of the variance of the intended outcome measure. If this estimate is imprecise it can impact on the power of the resulting trial. Therefore, in order to gain an accurate estimate it would be useful to have a similarly designed trial to aid in the estimation of the required parameters to be used in the design of the main trial. Pilot trials can be used to not only estimate the variance anticipated to be observed in the main trial but also to test the trial processes and procedures before launching in to the full scale main trial.

The estimate of the variability achieved from the pilot trial however is estimated with uncertainty. The imprecision in this estimate can impact on the sample size calculation (Kraemer et al., 2006). Two methods for adjusting the sample size calculation to allow for

this uncertainty have been described: the UCL approach (Browne, 1995) and the NCT approach (Julious and Owen, 2006).

Having insufficient power to detect a difference between treatments would be a potentially costly mistake for an investigator. In an attempt to lower the chance of an underpowered trial a sample size recalculation could be carried out at the end of an internal pilot trial. This allows the re-estimation of the variance from the actual trial population (Friede and Kieser, 2006). An initial proportion of the main trial data is collected and the sample size recalculated based on the new observed variance. The sample size is increased if the new recalculated sample size is larger than the original estimate, this is referred to as the restricted approach (Wittes and Brittain, 1990). The participants used in the sample size recalculation are also included in the final analysis of the trial.

These methods could impact on the power and required sample size of the main trial but also the required sample size of the internal pilot trial and the external pilot trial. Therefore, the aims of this thesis were to:

- Provide background information on the area of pilot trials, including definitions, current sample sizes and analysis methods (Section 8.2.1, 8.2.3 and 8.2.6),

- Investigate how using an estimate of the variance from a pilot trial (external and internal) to plan a main trial affects the power and sample size of the main trial (described in Section 8.2.2 and Section 8.2.4 below),

- Explore methods of setting a sample size for pilot trials (external and internal) which aim to minimise the overall trial sample size (outlined in Section 8.2.2 and Section 8.2.4 below) and,

- Examine how the relative cost of the external pilot trial versus the main trial affects the sample size of the two trials to minimise the overall trial cost (discussed in Section 8.2.3 below).

This chapter starts by summarising the chapters of this thesis (Section 8.2). Before going on to discuss in Section 8.3 the limitations of the work and suggestions for ideas of areas where further work could be carried out. Finally drawing conclusions from the work presented in Section 8.4.

8.2 Summary of Work

This section describes the work conducted and discusses the outcomes of each thesis chapter. Section 8.2.1 describes the background information gathered in the literature reviews and work presented in Chapter 1. Section 8.2.2 presents the work from Chapter 2 looking at the traditional methods for calculating the main trial sample size and the problems with these approaches. Section 8.2.3 summarises Chapter 3 on sample size justifications currently employed to choose a sample size for an external pilot trial. Section 8.2.4 condenses the work presented in Chapter 4 surrounding sample size requirements for an external pilot trial. Section 8.2.5 looks at Chapter 5 which extended this work to minimise the overall cost of a trial rather than the overall sample size. Section 8.2.6 reviews Chapter 6 which looked at internal pilot trial methodology and sample size recalculations. Finally 8.2.7 describes the effect that an internal pilot trial has and the power of a main trial before finally looking at what sample sizes may be required for internal pilot trials.

8.2.1 Background

Chapter 1 gives a brief description of the methodology behind clinical trials. A clinical trial that involves using a control treatment, either active or a placebo, as well as the treatment under investigation is considered to be controlled (Pocock, 1983). The trial is said to be

randomised if the type of treatment or the order of treatment which the participant receives is randomly allocated (Torgerson and Torgerson, 2008). A trial that has both a control group and involves randomisation is called a randomised controlled trial. The discussion in this chapter moves on to explaining how public clinical trials are funded in the UK. There are two main public funding bodies for health research in the UK, the Medical Research Council and the National Institute for Health Research.

The differences if any, between a pilot and a feasibility trial were discussed and the definition of a pilot trial used for this thesis was given in Section 1.4.2. The work in the thesis Section 1.4 on comparing the two terms pilot and feasibility has been published (Whitehead et al., 2014). The disagreement in the definitions of pilot and feasibility casues issues for researchers. Knowing what to call your project when designing a study can be difficult. The spurious naming of studies can make conducting audits of previous research difficult, searching through trials can be hard. Hence this can be misleading for researchers.

It my not be helpful to insist on specific terminology for preliminary studies, it is more important that the study has well defined suitable aims and objectives, is well designed with sufficient sample size to achieve its aims, has an appropriate analysis and is fully reported and published. However, the NIHR distinguish betweeen the terms, the NIHR definitions are becoming well used and followed which should be helpful in the future. There has recently been a push to improve the reporting of pilot and feasibility trials with the development of a CONSORT statement for the reporting of pilot and feasibility trials and a journal specifically for pilot and feasibility trials to allow these studies to be published.

Chapter 1 highlights the importance of sample size calculation in clincial trials. If a trial has too few participants the probability that the trial will find a statistically significant result even if one exists will be low (i.e. the power of the trial will be low). If the trial has too

many participants, resources are wasted, the treatment could have been shown to be inferior or superior with fewer participants (Altman, 1990).

The research results in Section 1.6 have been published (Billingham et al., 2013). This work looked at the current sample sizes of pilot trials registered on the UKCRN database. Of the 79 trials collected 21 were publicly funded pilot trials with a continuous endpoint. For these 21 trials the median sample size was 30 with an IQR of 20-60. The average sample size of 30 is as recommended by Browne (1995) and would be sufficient for any medium standardised effect size for the stepped rules of thumb presented in Chapter 4. The upper quartile of 60 is less than the 70 recommended by Teare et al. (2014) for precision around an estimate. The lower quartile of 20 matches the lowest recommendation seen throughout the thesis that means that in quarter of the trials have a sample size less than 20, this is worryingly small. A sample size this small would probably be insufficient to meet the aims and objectives of the study. In the audit 50% of the trials where larger than 30 participants these trials would be too large for standardised effect sizes above 0.3 according to the rules presented in Chapter 4.

Section 1.7 described a study that was carried out by a Wellcome Trust Summer Intern supervised during the thesis research, which looked at how predictive pilot trials are of main trials. It showed that in terms of the dropout of patients the bias is minimal however, the spread of the data is large. For the ratio of randomised to eligible patients less data was available again the bias was seen to be minimal with the main trials having a higher rate of converting eligible patients to randomised patients than the pilot trials. This may be expected as if the recruitment in to the trial was seen to be poor in the pilot trial then it is likely that the investigators would put procedures into place to try to remedy this in the main trial. This trend was seen in the paper by McDonald et al. (2006) where 53% of the trials which had a pilot trial in the review made changes to the recruitment strategy for the main trial.

Section 1.8 describes a piece of published work undertaken during the thesis (Lee et al., 2014) and presented at the Royal Statistical Society Conference 2014, which discusses how pilot trials are analysed. Pilot trials are usually not powered to detect a clinically relevant difference between the treatments under investigation and therefore it may be inappropriate to analyse a trial using the traditional hypothesis test and P-value. The paper recommends how confidence intervals of differing widths could be used to help display the strength of evidence from the pilot trial and the direction of the treatment effect. If this method of estimation is the primary outcome of the trial then the sample size should be set using the confidence interval approach, to provide sufficient sample size for this evaluation. This idea that pilot trials do not need to be powered in the traditional way is discussed further in Chapter 3 and carried through the thesis.

8.2.2 Main Trial Sample Size Calculations

Chapter 2 outlines the approaches to calculating a sample size for a main trial where the trial design is a superiority trial with independent groups and the outcome is Normally distributed. The sample size required is proportional to the variance of the outcome measure i.e. as the variance increases more participants are required. However, this variance must be estimated and the precision of this estimate may impact the power of the main trial. Methods for adjusting this estimate to preserve power in the main trial were discussed, the UCL approach (Browne, 1995) and the NCT approach (Julious and Owen, 2006). The UCL approach inflates the variance estimate to the $100X\%$ upper confidence limit to give a probability of X of achieving the required power. The NCT approach uses the fact that the variance is chi-squared distributed to choose a sample size for the main trial which will give the required power on average. The effect of both of these methods was studied further in Chapter 4.

As discussed in Chapter 1 traditional sample size calculations are based on the assumption that the analysis of the trial will be through hypothesis testing however, a pilot trial may not have an hypothesis testing objective of, say, looking for superiority of one treatment

over the other. Instead pilot trials are designed to estimate the variance or other parameters, and to test trial procedures. Pilot trials would therefore need a sample size justification which is based on its objectives and therefore, the sample size calculations presented in Chapter 2 may not always be appropriate when designing a pilot trial.

8.2.3 Pilot Trial Sample Size Justifications

Chapter 3 investigated methods for choosing a sample size for pilot trials that estimate the variance of the primary outcome to be used in the main trial. These could be precision based, based on maintaining the power in the main trial, proportional to the main trial sample size or described based on minimising the overall trial sample size of the pilot and the main trial together. This idea of minimising the overall trial sample size forms the basic idea, which is expanded in Chapters 4, 5 and 7. The effects of using the other suggested sample sizes are also studied in Chapter 4.

Ordinarily using hypothesis testing to test the efficacy of a treatment would not be recommended as an objective for a pilot or feasibility trial. This is usually saved for the main trial. However, these mat be circumstances in which hypothesis testing is used in a pilot or feasibility trial, probably based on a surrogate outcome measure for the true desired endpoint. For example, if using the true outcome would lead to a very large expensive trial, funders may want some indication that the treatment is having some effect i.e. a proof of concept.

The sample size of any study however, should enable you to achieve the aims of the trial and therefore if hypothesis testing is being carried out it is essential that a power calculation is carried out to calculate the sample size of the trial. This is to prevent small underpowered trials form being conducted which have limited scientific validity and are thought to be unethical, as discussed in Chapter 2.

8.2.4 Sample Sizes for External Pilot Trials to Minimise the Overall Trial Sample Size

If an adjustment method is used to allow for the imprecision in the variance estimate when the main trial sample size calculation is undertaken then increasing the sample size in the pilot trial to improve the precision of the variance estimate may not always be outweighed by the subsequent reduction in the sample size of the main trial.

Kieser and Wassmer (1996) and Sim and Lewis (2012) suggested external pilot trial sample sizes for minimising the overall trial sample size using the UCL approach, for limited values of the standardised effect size. However, in Chapter 4 the UCL approach was shown to be overly conservative leading to larger sample size than necessary compared to the NCT approach which considers the sampling distribution of the variance to allow for the imprecision of the estimate from a small pilot trial. Sample sizes which minimise the overall trial sample size for the NCT approach are smaller compared to those calculated relating to the UCL method.

Adjusting for the imprecision in the variance estimate helps to protect the trial from underpowering due to an inaccurate estimate of the variance. A small sample size for a pilot trial means that there will be a large amount of imprecision in the estimate of the variance and therefore when using an adjustment method to estimate the main trial sample size. Conversely, if the pilot trial sample size is large the estimate of the variance will be precise and the inflation to the main trial sample size will be small. However, as highlighted in Chapter 4 eventually the increase in the pilot trial sample size will not be offset by the subsequent reduction in the main trial sample size and therefore there is a pilot trial sample size which leads to a minimum overall trial sample size of the pilot and main trial added together.

The current methods for setting the pilot trial sample size are either based on fixed or proportional rules. A proportional rule means that the pilot trial size is proportionate to

the size of the main trial. The flat rules of thumb are fixed no matter the size of the main trial however, it was shown in Chapter 4 that in order to minimise the overall sample size of the trial, the larger the main trial the larger the pilot trial should be. Additionally for the proportional rules it was again shown in Chapter 4 that the optimal pilot trial sample size to minimise the overall trial sample size is not a specific proportion of the main trial and changes over the range of effect sizes discussed. Therefore, no one pilot trial sample size is optimal for all effect sizes.

In Section 4.2.2 methods for estimating the minimum overall sample sizes using the NCT approach were proposed. However, using these values would require the investigator to know the required effect size for the main trial before the pilot trial. As this perhaps may be unrealistic in some cases further results were proposed where stepped rules of thumb for the pilot sample size were introduced. These steps were based on bands for the treatment differences of extra small, small, medium and large standardised effect sizes (see Table 8.1).

Table 8.1: Stepped Rules of Thumb for the NCT Approach Sample Sizes are for a Two-armed Trial

Standardised Effect Size	80% Powered Main Trial	90% Powered Main Trial
Extra Small ($\delta < 0.1$)	100	150
Small ($0.1 \leq \delta < 0.3$)	40	50
Medium ($0.3 \leq \delta < 0.7$)	20	30
Large ($\delta \geq 0.7$)	20	20

By carrying out the method proposed and minimising the required overall trial sample size we could reduce the sample size needed to run clinical trials, reduce the cost of trials, reduce the trial duration or perhaps increase the feasibility of trials involving populations/ conditions where numbers are restricted e.g.in rare conditions. The work presented in Chapter 4 is published (Whitehead et al., 2015) and has been presented at Statistics Research Student Conferences and the Society for Clinical Trials Conference.

8.2.5 Sample Sizes for External Pilot Trials to Minimise the Overall Trial Cost

Minimising the overall trial sample size does not only have ethical advantages for the numbers of patients used, but also for the cost of trials. However, depending on the relative costs between the pilot and the main trial minimising the sample size may not necessarily minimise the overall cost of the trial. Therefore, the pilot trial sample sizes which lead to the overall minimum trial sample size may not lead to the overall minimum cost for the trial. Chapter 5 looked at how the balance of the costs between the two trials affects the sample sizes which result in the minimum cost of the trial overall.

Table 8.2 gives the proposed pilot sample sizes for the same banded effect sizes given in Table 8.1. It can be seen that the optimal sample sizes change as the relative cost (R) moves away from a one to one ratio. This reflects the fact that as it becomes cheaper to enter a participant into the pilot it is cheaper overall to have a larger pilot and increase the precision of estimates for the main trial sample size. Alternatively if the main trial is less expensive than the pilot it may be less costly overall to accept the imprecision from a smaller pilot trial and have a relatively large main trial. The most likely scemario is that R will be larger than 1. There is likely to be an amount of fixed cost involved with every trial. For a pilot trial with less patients this fixed cost per patient is likely to be high. For a main trial with a larger sample size the fixed cost per patient is likely to be lower, therefore leading to an R value greater than 1. It should be noted that in Table 8.2 for R = 1 the results are the same as in Table 8.1.

Table 8.2: Pilot Trial Sample Sizes for Varying Relative Cost of the Pilot and Main Trial for a Two-armed Trial

Standardised Effect Size	Relative Cost	80% Powered Main Trial	90% Powered Main Trial
Extra Small	$R < 1$	240	260
	$R = 1$	100	150
	$1 < R \leq 5$	90	140
	$5 < R \leq 20$	50	60
	$R > 20$	30	40
Small	$R < 1$	60	80
	$R = 1$	40	50
	$1 < R \leq 5$	30	40
	$5 < R \leq 20$	20	20
	$R > 20$	20	20
Medium	$R < 1$	30	40
	$R = 1$	20	30
	$1 < R \leq 5$	20	20
	$5 < R \leq 20$	20	20
	$R > 20$	20	20
Large	$R < 1$	20	30
	$R = 1$	20	20
	$1 < R \leq 5$	20	20
	$5 < R \leq 20$	20	20
	$R > 20$	20	20

8.2.6 Internal Pilot Trials and Sample Size Recalculations

An internal pilot trial is a pilot trial were the sample forms the first part of the main RCT and the participants contribute to the final analysis (NETSCC, 2012). A sample size recalculation can be carried out at the end of an internal pilot trial to re-estimate the variance estimate for the sample size calculation based on the real trial data. The sample

size recalculation can be done in a restricted or unrestricted manner. The restricted approach is most common in publicly funded research (Dimairo et al., 2015). Although the internal pilot trial design protects against under-powering. This restricted approach can lead to overpowering if the initial estimate of the variance is too high. Again a minimum sample size for an internal pilot trial is suggested to be 20 participants. However, most of the sample size recommendations for internal pilot trials are proportional rules (Wittes and Brittain, 1990, Wittes at al., 1999). The initial stages of this work have been presented at the Statistics Research Students Conference and the Royal Statistical Society Conference 2015.

Birkett and Day (1994) suggested only setting the size of the internal pilot trial at the start so that the probability of the trial being larger than necessary would be significantly reduced. However, this in reality is may be impractical in a public funded setting where the funders are likely to require some estimate of the trials size and duration, which would additionally impact on the budgeting and cost of the trial. Therefore the sample size of the internal pilot trial is a point of interest.

Other than the 10 per arm sample size suggested as a minimum by Birkett and Day (1994) several suggestions on internal piot trial sample size have been made, from a quarter to three quarters of the planned main trial sample size. A balance needs to be made between getting an account estimate of the variance versus conducting the sample size recalculation early enough to allow the investigator to plan/run the study effectively.

If the sample size recalculation is early in the trial the sample size can be adjusted if necessary with lots of warning before the extra patients are need to be recruited giving the time for the nescessary administrative procedures and funding extensions to be in place. However, if this happens early you sacrifice the accuracy of the estimate of the variance.

Other issues could also affect the decision of what sample size to use for the internal pilot trial: the size if the trial, the length of the anticipated recruitment period during the trial or the length of time between the intervention and the collection of the outcome data. For example, if the recruitment window is short and the follow up time to the data collection is long, waiting until a large amount of participants have completed follow up could mean that you have already recruited everyone into your trial before your sample size recalculation, and reopening recruitment could be problematic.

8.2.7 The Effect of an Internal Pilot Trial on the Main Trial Power and Required Sample Sizes

An internal pilot trial design protects against underpowering when the original estimate of the variance is too low when compared to the variance seen in the actual main trial. However, in Chapter 7 the restricted design was shown to lead to overpowering if the variance estimate was initially too high. Therefore, on average the internal pilot trial design leads to achieving a higher power than initally planned. Consequently, on average it was shown in Chapter 7 that the restricted design also requires more partcipants than the fixed design. However, it was highlighted how altering the nominal or recalculation power could combat this effect.

It was shown in Chapter 7 that to get the optimal estimate of the sample size required we need to include every participant from the trial in the internal pilot trial to get the most accurate prediction of the variance i.e. wait until the end of the trial when we have all patients recruited and followed up to estimate the variance. However, this is not probably practical in reality. What needs to be balanced is the need to get an accurate estimate of the variance, while performing the internal pilot trial early enough in the trial for it to be useful to the investigators and the funders. Suggested sample sizes given in Chapter 7 which may provide an accurate estimate of the variance from an internal pilot trial are presented in Table 8.3.

Table 8.3: Sample Size Recommendations for Internal Pilot Trials for a Two-armed Trial

Standardised Effect Size	80% Powered Trial	90% Powered Trial
Extra Small ($\delta < 0.1$)	160	190
Small ($0.1 \leq \delta < 0.3$)	30	40
Medium ($0.3 \leq \delta < 0.7$)	20	20
Large ($\delta \geq 0.7$)	20	20

The sample size of the internal pilot was fixed in Section 7.5 to allow the investigation of the properties of the designs and the estimation of the effects of the sample size on the expected power and sample size of the main trial. Using a proportional rule to choose the pilot trial sample size would mean that the variance estimates at end of the internal pilot is dependent on the variance estimate from the external pilot. The effect of this on the distributions of the variances was not investigated in this thesis, but could be examined further in future work

8.3 Limitations and Areas for Further Work

This section outlines some limitations of the work performed in this thesis and aims to describe some suggestions for further work, which could arise from these. Section 8.3.1 discusses using alternative endpoints for the trials outcome measure or aims of the trial. Section 8.3.2 considers ideas for ways to improve the estimate of the variance by combining data from several trials. Section 8.3.3 describes how adaptive trial designs might affect the results given in this thesis. Section 8.3.4 reviews the idea of placing bounds on the possible alterations to the sample size at the interim and Section 8.3.5 considers the way in which dropouts are allowed for.

8.3.1 Alternative Endpoints or Aims

Although this thesis set out to establish the required sample size for pilot trials, it concentrated on continuous Normally distributed outcome measures. However, trials

may have endpoints which are binary, ordinal or survival outcomes; as discussed in Chapter 2.

The sample size requirement for a binary endpoint is calculated using the approximate formula,

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \, (p_A(1 - p_A) + p_B(1 - p_B))}{\delta^2}, \qquad (8.1)$$

this can be used if p_A and p_B are larger than 0.05 (Campbell et al., 1995). Where p_A is the proportion of patients in treatment group A that have the event and p_B is the proportion in treatment group B. This thesis concentrates on accurately estimating the variance of the outcome measure for a continuous outcome variable, however, the calculation for a binary outcome requires different parameters to be estimated. For example, the event rate in the control group. Further work could investigate how capable pilot trials are at giving a good prediction of the event rate in the control group and what affects their ability to predict it. Perhaps looking to recommend sample sizes for pilot trials with a binary endpoint which, in Chapter 1 were shown to be currently larger on average (36 versus 30) compared to pilot trials with continous endpoints (Billingham et al., 2013).

Similarly, the investigations could extend to ordinal or survival data. The sample size requirement for ordinal/ categorical outcomes also requires the estimation of the odds ratio of a patient being in each category or less compared to the others (Campbell et al., 1995). Estimating the sample size needed for survival endpoint is a little different from the other types of calculation but involves specifying a clinically relevant hazard ratio for the event between the groups and estimating the rate of events in each group (Machin et al., 2009).

For the work in this thesis to be able to be extended to other distributional forms then these distributions would need to have a Normal approximation which may not always be the case for small pilot studies.

Where the work could be potentially be extended is to trials with objectives other than superiority, as was a restriction in the thesis. Trials with the objective to show non-inferiority, equivalence trials and/or bioequivalence where the outcome is anticipated to take a Normal form could be an extension to the work. In an equivalence trial the null hypothesis is that the treatments have different effects i.e. $\mu_A \neq \mu_B$. In an equivalence trial the aim of the trial is the opposite to a superioirity trial now instead of wishing to prove that the teo treatments are different in terms of their effect on the outcome measure, we want to prove that they are the same. An estimate for the required sample size for this kind of trial can be calculated from the following formula (Julious, 2009),

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\alpha}\right)^2}{\left((\mu_A - \mu_B) - d\right)^2}.$$

Where μ_A is the expected mean in treatment group A, μ_B is the espected mean in treatment group B, the largest clinically acceptable effect for which equivalence can be declared is given by d and the population variance σ^2 is estimated in the same way as discussed thoroughout this thesis. The equation presented here is the equivalent of Equation 2.14 but for equivalence trials.

Moreover, the pilot trial could use a surrogate endpoint for the full clinical outcome measure to reduce the duration of the trial. In such a case it could be studied not only how the surrogate outcome predicts the clinical outcome, but also how the variance of the surrogate outcome predicts the variance of the clinical outcome if at all. Therefore investigating how the sample size calculation for the main trial could be based on predictions from a pilot trial which uses a surrogate endpoint. This would however, vary between disease area and outcome measures and so may be difficult to explore.

In this work only trials where the randomisation is between individual participants have been discussed but there may be occasions when this is not possible or logistically practical (Pocock, 1983). In these situations cluster randomisation might be possible where groups of people for example; a school or clinic, or a person responsible for a group of participants for example; a physiotherapist or a surgeon, are randomised to the interventions rather than the individuals (Campbell and Walters, 2014, Eldridge and Kerry, 2012).

For cluster randomised trials the sample size calculation is adjusted by,

$$1 + (n' - 1) \times ICC, \tag{8.2}$$

where n' is the average cluster size and the ICC is the intraclass correlation. The ICC is a measure of how similar to each other participants are within a cluster.

8.3.2 Combining Variance Estimates

The main aim of an internal pilot trial is to protect against underpowering in the main trial therefore, the restricted procedure which is the most common method allows us to re-estimate the variance mid-trial and adjust the sample size upwards if the re-estimated variance is higher than the original estimate. In a way this procedure replaces the original estimate with the new value. This could be regarded as a waste of information, especially if work has been carried out beforehand, for example, in cases where there is an external and an internal pilot trial.

Ways in which to combine variance estimates from previous trials with the current estimate could be investigated and their effect on the power and sample size required for the main and overall trial. This could involve taking weighted average of the variances, or perhaps weighting the estimate towards the more recent trials or the internal pilot trial.

Alternatively, the estimate of the variance could be updated using a Bayesian methodology to combine the prior information we might have with the internal pilot data, furthermore the Bayesian method could also be weighted (De Santis, 2006).

8.3.3 Adaptive Designs

In this thesis the only adaptive feature of a trial which has been investigated is a sample size recalculation at an interim. There are many other adaptations including early stopping for futility or superioirity of the experimental treatment, which could be applied in a trial as discussed in Chapter 6, Section 6.1.

It could also be interesting to extend the work in the thesis to look at the effect of the methods proposed in this thesis if for example, the promising zone approach (Mehta and Pocock, 2011) was undertaken or perhaps early stopping was allowed for at the interim time point. If early stopping was allowed for using an interim efficacy analysis of the treatment or the promising zone approach this may reduce the overpowering effect of the internal pilot trial design. The method used in Chapter 7 to look at the properties of the internal pilot trial design could be extended to investigate the promising zone approach by including another decision point in the process for estimating the trial sample size based on the pilot trial data, where it is also possible to stop the trial as well as continue as planned or increase the sample size. The effect of these approaches may also change the recommended sample sizes for the internal pilot trial design.

8.3.4 Sample Size Recalculation within Bounds

The standard internal pilot trial method allows the sample size to be readjusted upwards indefinitely and also when an increase in patient numbers is inconsequential for the trial as a whole or would result in minimal loss of power.

A simpler adjustment could be made to the sample size at the recalculation. Gould and Shih (1992) suggested a slight adaptation of the method so that the sample size requirement is only increased to the recalculated figure if it is some factor (f) bigger than the original to make the extension worthwhile,

$$\text{if} \frac{\mathrm{N}_{RECALC}}{\mathrm{N}_0} > f, \mathrm{N}_1 = \mathrm{N}_{RECALC} \tag{8.3}$$

$$\text{else } \mathrm{N}_1 = \mathrm{N}_0$$

they suggest a value for f of 1.25 in which case the sample size will only be increased if the new sample size is more than 25% higher than the initial sample size calculation requirement. In addition, they propose putting an upper limit on the re-estimated sample size of some function of the original sample size, for example, twice the originally planned sample size. This cap could be derived arbitrarily or be financially driven in that there is an upper limit on how many patients could be recruited due to costs. The idea is to keep the sample size of the trial within reasonable limits of what was originally planned. If the re-estimated sample size requirement is higher than this upper bound for the sample size the trial could either; continue to the upper bound at which the trial would cease and be analysed despite the possibility of lower than required power, or the trial could stop after the interim and its results summarised without hypothesis testing (Gould and Shih, 1992).

The lower limit for f also has an intuitive appeal. Operationally it may not be worth increasing a sample size for a nominal increase in the sample size and the study team may prefer a small loss in power to increasing the sample size by a small amount.

The bounded approach could be applied within the methods evaluated in this thesis. The technique would reduce the possibility of large as well as slight overpowering, changing the average sample sizes and average powers of the internal pilot trial design as presented in Chapter 7.

8.3.5 Accounting for Dropout

A major limitation of sample size calculations is that after the complicated statistical procedures are employed to get the best estimate of the required sample size of evaluable participants to achieve the specified power for a certain level of Type I error and MCID, the number is inflated by an estimate of the dropout rate to give the required number of people which need to be recruited into the trial. The number of people approached to be involved in the trial must also be larger than this new sample size as not everyone who is eligible and approached will end up being randomised in to the trial. Further work could involve looking in more detail at predicting the dropout from and recruitment in to trials to improve the accuracy of these inflations to the sample size.

8.4 Conclusion

It was shown in this thesis how if an adjustment method is to be used at the main trial sample size calculation to allow for the imprecision involved in estimating the variance, then the sample sizes recommended in the literature are not always optimal when it comes to minimising the overall trial sample size of the pilot and the main trial together, depending on the minimum clinically inportant difference. Therefore when planning a trial thought should be given to the effect of the chosen pilot trial sample size on the sample size of the subsequent randomised controlled trial. The investigator should consider whether an alternative pilot trial sample size would be more efficient overall in terms of the combined sample size of the pilot and main trials.

Sample size recommendations for external and internal pilot trials are made which aim to minimise the overall sample size of the trial. It was shown that the optimal pilot trial sample size increases with the size of the main trial and solutions were described to minimise the overall sample size in this context. To help in the applicability of the results

in practice, stepped rules of thumb for the pilot sample sizes were introduced which vary depending on the main trial sample size.

Sample sizes for external pilot trials which minimise the overall cost of the trial programme were also given. The results generated in this thesis show that when the pilot trial is less expensive per patient than the main trial the optimal pilot trial sample size increases, giving more precision for the variance estimate and a relatively small main trial. Conversely, when the pilot trial is more expensive per patient than the main trial the optimal pilot trial sample size decreases, accepting less precision from the pilot and thus a relatively larger main trial. Therefore when planning a trial thought should be given to the effect of the chosen pilot trial sample size on the cost of the subsequent randomised controlled trial. Using the results presented in this thesis the investigator should consider whether an alternative pilot trial sample size would be more efficient overall in terms of the combined cost of the pilot and main trials together.

Ideally we would gain as accurate estimate of the variance as possible from the pilot trial. However, the NHS is an area of limited resources (in terms of patients, money and staff). Investigators should make the best use of the resources allocated to them as possible. A lack of accuracy in estimating the variance from a pilot trial (small sample size) leads to a larger required sample size in the main trial. Conversely, a large amount of accuracy in the estimation of the variance from a pilot (large sample size) leads to a smaller required sample size in the main trial. There is a balance to be made, this thesis examines the cost/ benefit of this trade off of sample sizes between the two trials. Hence presenting a method of optimising the number of patients or financial costs used in a trial therefore maximising the utility of NHS resources

It is intended that the work in this thesis will help researchers planning and designing publicly funded clinical trials to justify their choice of sample size for pilot trials and to think about the effect the methods have on the power and required sample size of their main randomised controlled trial.

References

ABRAMOWITZ , M. & STEGUN, I. A. 1965. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* New York, Dover.

ALTMAN, D. A. 1980. Medicine and Mathematics - Statistics and Ethics in Medical Research III: How large a sample? *BMJ,* 281**,** 1336-1338.

ALTMAN, D. G. 1990. *Practical Statistics for Medical Research,* London, Boca Raton: Chapman and Hall/CRC.

ARAIN, M., CAMPBELL, M. J., COOPER, C. L. & LANCASTER, G. A. 2010. What is a Pilot or Feasibility Study? A Review of Current Practice and Editorial Policy. *BMC Medical Research Methodology,* 10**,** 67.

ARNOLD, D. M., BURNS, K. E. A., ADHIKARI, N. K. J., KHO, M. E., MEADE, M. O., COOK, D. J. & MCMASTER CRITICAL CARE, I. 2009. The Design and Interpretation of Pilot Trials in Clinical Research in Critical Care. *Critical Care Medicine,* 37**,** S69-S74.

BAUER, P. & KOHNE, K. 1994. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics,* 50**,** 1029-1041.

BILLINGHAM, S. A. M., WHITEHEAD, A. L. & JULIOUS, S. A. 2013. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Med Res Methodol,* 13**,** 104.

BIRKETT, M. A. & DAY, S. J. 1994. Internal Pilot Studies for Estimating Sample Size. *Statistics in Medicine,* 13**,** 2455-2463.

BLAND, M. 2000. *An Introduction to Medical Statistics,* New York, Oxford University Press.

BMJ. 2012. *Article Requirements* [Online]. Available: http://www.bmj.com/about-bmj/resources-authors/article-submission/article-requirements [Accessed 16th October 2012].

BROWNE, R. H. 1995. On the Use of a Pilot Sample for Sample Size Determination. *Statistics in Medicine,* 14**,** 1933-1940.

CAMPBELL, M., FITZPATRICK, R., HAINES, A., KINMONTH, A. L., SANDERCOCK, P., SPIEGELHALTER, D. & TYRER, P. 2000. Framework for Design and Evaluation of Complex Interventions to Improve Health. *BMJ,* 321**,** 694-696.

CAMPBELL, M. & WALTERS, S. 2014. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Services Research*, John Wiley and Sons Ltd.

CAMPBELL, M. J., JULIOUS, S. A. & ALTMAN, D. G. 1995. Estimating Sample Sizes for Binary, Ordered Categorical, and Continuous Outcomes in Two Group Comparisons. *BMJ,* 311**,** 1145-8.

CAMPBELL, M. J., MACHIN, D. & WALTERS, S. J. 2010. *Medical Statistics: A textbook for the health sciences*, John Wiley & Sons Inc.

CHALMERS, T. C. 1983. The Control of Bias in Clinical Trials. *In:* SHAPIRO, S. H. & LOUIS, T. A. (eds.) *Clinical Trials: Issues and Approaches.* New York: Marcel Dekker.

CHOW, S. C., WANG, H. & SHAO, J. 2003. *Sample Size Calculations in Clinical Research*, CRC Press.

CHUANG-STEIN, C., ANDERSON, K., GALLO, P. & COLLINS, S. 2006. Sample Size Re-estimation: A Review and Recommendations. *Drug Information Journal,* 40**,** 475-484.

COCKS, K. & TORGERSON, D. J. 2013. Sample size calculations for pilot randomized trials: a confidence interval approach. *Journal of Clinical Epidemiology,* 66**,** 197-201.

COHEN, J. 1992. A Power Primer. *Psychological Bulletin,* 112**,** 155-159.

CONSORT. 2012. *CONSORT 2010 Checklist* [Online]. Available: http://www.consort-statement.org/consort-statement/overview0/#checklist [Accessed 16th October 2012].

COOKSEY, S. D. 2006. A Review of UK Health Research Funding. *In:* TREASURY, H. (ed.).

COOPER, C. L., HIND, D., PARRY, G. D., ISAAC, C. L., DIMAIRO, M., O'CATHAIN, A., ROSE, A., FREEMAN, J. V., MARTIN, L., KALTENTHALER, E. C., THAKE, A. & SHARRACK, B. 2011. Computerised Cognitive Behavioural Therapy for the Treatment of Depression in People with Multiple Sclerosis: External Pilot Trial. *Trials,* 12**,** 259.

CRAWLEY, M. J. 2015. *Statistics. An Introduction using R,* Chichester, Wiley.

CTRU, T. U. O. S. 2016. *Big CACTUS* [Online]. Available: https://www.sheffield.ac.uk/scharr/sections/dts/ctru/bigcactus [Accessed 25th January 2016].

DALGAARD, P. 2008. *Introductory Statistics with R,* New York, Springer.

DALY, L. E. & BOURKE, G. J. 2008. *Interpretation and Uses of Medical Statistics*, Wiley-Blackwell.

DE SANTIS, F. 2006. Power Priors and Their Use in Clinical Trials. *The American Statistician,* 60**,** 122-129.

DEGRUTTOLA, V., CLAX, P., DEMETS, D., DOWNING, G., ELLENBERG, S., FREDMAN, L., GAIL, M., PRENTICE, R., WITTES, J. & ZEGER, S. 2001. Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institute of Health Workshop. *Controlled Clinical Trials* 485 - 502.

DENNE, J. S. & JENNISON, C. 1999. Estimating the Sample Size for a T-Test Using an Internal Pilot. *Statistics in Medicine,* 18**,** 1575-1585.

DIEM, K. 1962. *Documenta Geigy: Scientific Tables,* Manchester, Geigy Pharmaceutical Company Limited.

DIMAIRO, M., JULIOUS, S., TODD, S., NICHOLL, J. & BOOTE, J. 2015. Cross-sector surveys assessing perceptions of key stakeholders towards barriers, concerns and facilitators to the appropriate use of adaptive designs in confirmatory trials. *Trials,* 16**,** 585.

DOBSON, A. J. 2001. *An Introduction to Generalized Linear Models,* Boca Raton, Chapman & Hall/CRC.

DU PREL, J. B., HOMMEL, G., RÖHRIG, B. & BLETTNER, M. 2009. Confidence Interval or P-Value?: Part 4 of a Series on Evaluation of Scientific Publications. *Dtsch Arztebl Int,* 106**,** 335-9.

ELDRIDGE, S. & KERRY, S. 2012. *A Practical Guide to Cluster Randomised Trials in Health Services Research,* Chicester, John Wiley & Sons Ltd.

ELDRIDGE, S. , LANCASTER, G. A., CAMPBELL, M. J., THABANE L., HOPEWELL, S., COLEMAN, C. L. & BOND, C. M. 2016. Defining Feasibility and Pilot Studies in Preparation for

Randomised Controlled Trials: Development of a Conceptual Framework. PLoS ONE 11(3): e0150205. doi:10.1371/journal.pone.0150205

EMEA 2007. Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design.

EVANS, D. 2003. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing,* 12**,** 77-84.

FLEISS, J. L. 1986. *The Design and Analysis of Clinical Experiments,* New York, Wiley.

FRIEDE, T. & KIESER, M. 2001. A Comparison of Methods for Adaptive Sample Size Adjustment. *Statistics in Medicine,* 20**,** 3861-3873.

FRIEDE, T. & KIESER, M. 2006. Sample Size Recalculation in Internal Pilot Study Designs: A Review. *Biometrical Journal,* 48**,** 537-555.

FRIEDMAN, L. M., FURBERG, C. D. & DEMETS, D. L. 2010. *Fundamental of Clinical Trials,* New York, Springer Science and Business Media.

GALLO, P., CHUANG-STEIN, C., DRAGALIN, V., GAYDOS, B., KRAMS, M. & PINHEIRO, J. 2006. Adaptive designs in clinical drug development--an Executive Summary of the PhRMA Working Group. *J Biopharm Stat,* 16**,** 275-83; discussion 285-91, 293-8, 311-2.

GONZALEZ, C. D., BOLAÑOS, R. & DE SEREDAY, M. 2009. Editorial on hypothesis and objectives in clinical trials: superiority, equivalence and non-inferiority. *Thromb J,* 7**,** 3.

GOODACRE, S. W., BRADBURN, M., CROSS, E., COLLINSON, P., GRAY, A. & HALL, A. S. 2010. The Randomised Assessemnt of Treatment using Panel Assay of Cardiac Markers (RATPAC) trial: A Randomised Controlled Trial of Point-Of-Care Cardiac Markers in the Emergency Department. *Heart Online*.

GOULD, A. L. & SHIH, W. J. 1992. Sample Size Re-estimation Without Unblinding for Normally Distributed Outcomes with Unknown Variance. *Communications in Statistics - Theory and Methods,* 21**,** 2833-2853.

HALPERN, S. D., KARLAWISH, J. H. T. & BERLIN, J. A. 2002. The Continuing Unethical Conduct of Underpowered Clinical Trials. *JAMA,* 288.

HERTZOG, M. A. 2008. Considerations in Determining Sample Size for Pilot Studies. *Research in Nursing & Health,* 31**,** 180-191.

HIND, D., SCOTT, E. J., COPELAND, R., BRECKON, J. D., CRANK, H., WALTERS, S. J., BRAZIER, J. E., NICHOLL, J., COOPER, C. & GOYDER, E. 2010. A Randomised Controlled Trial and Cost-Effectiveness Evaluation of 'Booster' Interventions to Sustain Increases in Physical Activity in Middle-Aged Adults in Deeprived Urban Neighbourhoods. *BMC Public Health,* 10.

HIORNS, R. W. 1971. *Statistics Definitions and Formulae for Students*, Sir Isaac Pitman and Sons Ltd.

HTA. 2012. *Guidance Notes for Completing the Online Evidence Synthesis Full Preposal Form* [Online]. Available: http://www.hta.ac.uk/funding/clinicaltrials/Feb_2012_HTA_CET_Guidance_EVIDENCE_SYNTHESIS_FULL_form.pdf [Accessed 16th October 2012].

ICH. 1998. *E9: Statistical Principals for Clinical Trials* [Online]. Available: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf [Accessed 18th October 2012].

JAESCHKE, R., SINGER, J. & GUYATT, G. H. 1989. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials,* 10**,** 407-15.

JENNISON, C. & TURNBULL, B. W. 1999. *Group Sequential Methods with Applications to Clinical Trials*, CRC Press.

JULIOUS, S. A. 2004. Designing Clinical Trials with Uncertain Estimates of Variability. *Pharmaceutical Statistics,* 3**,** 261-268.

JULIOUS, S. A. 2005. Sample Size of 12 per Group Rule of Thumb for a Pilot Study. *Pharmaceutical Statistics,* 4**,** 287-291.

JULIOUS, S. A. 2009. *Sample Sizes for Clinical Trials*, Chapman and Hall.

JULIOUS, S. A. & OWEN, R. J. 2006. Sample Size Calculations for Clinical Studies Allowing for Uncertainty about the Variance. *Pharmaceutical Statistics,* 5**,** 29-37.

JULIOUS, S. A. & PATTERSON, S. D. 2004. Sample Sizes for Estimation in Clinical Research. *Pharmaceutical Statistics,* 3**,** 213-215.

JULIOUS, S. A., TAN, S. B. & MACHIN, D. 2010. *An Introduction to Statistics in Early Phase Trials*, John Wiley & Sons, Ltd.

KAIRALLA, J., COFFEY, C., THOMANN, M. & MULLER, K. 2012. Adaptive trial designs: a review of barriers and opportunities. *Trials,* 13**,** 145.

KIANIFARD, F. & ISLAM, M. Z. 2011. A Guide to the Design and Analysis of Small Clinical Studies. *Pharmaceutical Statistics,* 10**,** 363-368.

KIESER, M. & FRIEDE, T. 2000. Re-calculating the Sample Size in Internal Pilot Study Designs with Control of the Type I Error Rate. *Statistics in Medicine,* 19**,** 901-911.

KIESER, M. & WASSMER, G. 1996. On the Use of the Upper Confidence Limit for the Variance from a Pilot Sample for Sample Size Determination. *Biometrical Journal,* 8**,** 941-949.

KIRKWOOD, B. R. & STERNE, J. A. C. 2003. *Essential Medical Statistics*, Blackwell Science Ltd.

KNOX, C., WALTERS, S. J., HIND, D. & JULIOUS, S. A. 2014. Audit of Recruitment and Retention to Publicly funded Randomised Controlled Trial (RCTs). *In:* SHEFFIELD, U. O. (ed.) *RSS Conference Poster.*

KRAEMER, H. C., MINTZ, J., NODA, A., TINKLENBERG, J. & YESAVAGE, J. A. 2006. Caution Regarding the Use of Pilot Studies to Guide Power Calculations for Study Proposals. *Arch Gen Psychiatry,* 63.

LANCASTER, G. A., DODD, S. & WILLIAMSON, P. R. 2004. Design and Analysis of Pilot Studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice,* 10**,** 307-312.

LANCET. 2012. *Types of Article and Manuscript Requirements* [Online]. Available: http://www.thelancet.com/lancet-information-for-authors/article-types-manuscript-requirements [Accessed 16th October 2012].

LEE, E. C., WHITEHEAD, A. L., JACQUES, R. M. & JULIOUS, S. A. 2014. The Statistical Interpretation of Pilot Trials: Should significance thresholds be reconsidered? *BMC Medical Research Methodology,* 14**,** 41-41.

LESAFFRE, E. 2008. Superiority, Equivalence and Non-Inferiority Trials. */bukketin of the NYU Hospital for Joint Diseases,* 66**,** 150-154.

MACHIN, D., CAMPBELL, M. J. & TAN, S.-B. 2009. *Sample Size Tables for Clinical Studies*, BMJ Books.

MACHIN, D., CAMPBELL, M. J., TAN, S. B. & TAN, S. H. 2008. *Sample Size Tables for Clinical Studies,* Chichester, Wiley-Blackwell.

MCDONALD, A. M., KNIGHT, R. C., CAMPBELL, M. K., ENTWISTLE, V. A., GRANT, A. M., COOK, J. A., ELBOURNE, D. R., FRANCIS, D., GARCIA, J., ROBERTS, I. & SNOWDON, C. 2006. What Influences Recruitment to Randomised Controlled Trials? A Review of Trials Funded by Two UK Funding Agencies. *Trials,* 7.

MEHTA, C. R. & POCOCK, S. J. 2011. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine,* 30**,** 3267-3284.

MOSHMAN, J. 1958. A Method for Selecting the Size of the Initial Sample in Stein's Two Sample Procedure. *The Annals of Mathematical Statistics,* 29**,** 1271-1275.

MRC 2000. A Framework for the Development and Evaluation of RCTs for Complex Interventions to Improve Health. *London: MRC*.

MRC. 2012a. *About Us: Facts and Figures* [Online]. Available: http://www.mrc.ac.uk/About/Factsfigures/index.htm [Accessed 10th October 2012].

MRC. 2012b. *About Us: Mission Statement* [Online]. Available: http://www.mrc.ac.uk/About/Missionstatement/index.htm [Accessed 10th October 2012].

MRC. 2012c. *Our Research: In Practice* [Online]. Available: http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/PHSPolicy/inpractice/index.htm [Accessed 16th October 2012].

NETSCC. 2012. *Glossary: Feasibility and Pilot Studies* [Online]. Available: http://www.netscc.ac.uk/glossary/ [Accessed 8th October 2012].

NICE. 2013. *Glossary* [Online]. Available: http://www.nice.org.uk/website/glossary/glossary.jsp?alpha=P [Accessed 30th July 2013].

NIGMS. 2011. *Evaluation A-Z Glossary* [Online]. Available: http://www.nigms.nih.gov/nigms.nih.gov/Templates/CommonPage.aspx?NRMODE=Published&NRNODEGUID={EEAC2F09-234D-44DA-BA80-B32CE242FCA0}&NRORIGINALURL=%2fResearch%2fEvaluation%2fglossary.htm&NRCACHEHINT=Guest#P [Accessed 27 September 2012].

NIGMS. 2012. *About NIGMS* [Online]. Available: http://www.nigms.nih.gov/About/ [Accessed 27 September 2012].

NIHR 2012a. Funding Oppurtunities for Research and Career Development.

NIHR 2012d. NETSCC: Needs-led and Science-added Management of Evalaution Research on behalf of the National Institute for Health Research.

NRES. 2012. *Integrated Research Application System* [Online]. Available: https://www.myresearchproject.org.uk/help/IRASIndex.aspx [Accessed 29th October 2012].

O'BRIEN, P. C. & FLEMING, T. R. 1979. A multiple testing procedure for clinical trials. *Biometrics,* 35**,** 549-56.

PALMER, R. 2015. *A Study to Assess the Clinical and Cost Effectiveness of Aphasia Computer Treatment Versus Usual Stimulation or Attention Control Long Term Post Stroke (CACTUS) Research Protocol (Version 3.0)* [Online]. Available: https://www.sheffield.ac.uk/polopoly_fs/1.477276!/file/BigCACTUSProtocolv3.0_12Feb15.pdf [Accessed 25th January 2016].

PALMER, R., ENDERBY, P., COOPER, C., JULIOUS, S., DIXON, S., MORTLEY, J., PATERSON, G. & LATIMER, N. 2011. Cost Effectiveness of Aphasia Computer Therapy versus Usual Stimulation: A Pilot Randomised Controlled Trial (CACTUS). Sheffield Teaching Hospitals NHS Foundation Trust.

PARKER, R. A. & BERMAN, N. G. 2003. Sample Size: More than Calculations. *The American Statistician,* 57**,** 166-170.

PARSONS, H. M. 1974. What Happened at Hawthorne? *Science,* 183**,** 922-932.

PETRIE, A. & SABIN, C. 2013. Medical Statistics at a Glance. Wiley - Blackwell.

POCOCK, S. J. 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika,* 64**,** 191-199.

POCOCK, S. J. 1983. *Clinical Trials: A Practical Approach*, John Wiley & Sons Ltd.

POSCH, M., BAUER, P. & BRANNATH, W. 2003. Issues in designing flexible trials. *Statistics in Medicine,* 22**,** 953-969.

PRENTICE, R. L. 1989. Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria. *Statistics in Medicine,* 8**,** 431-440.

PROSCHAN, M., LIU, Q. & HUNSBERGER, S. A. 2003. Practical Midcourse Sample Size Modification in Clinical Trials. *Controlled Clinical Trials,* 24**,** 4-15.

PUFFER, S. & TORGERSON, D. 2003. Recruitment Difficulties in Randomised Controlled Trials. *Controlled Clinical Trials***,** 214-15.

RAI, S. K., YAZDANY, J., FORTIN, P. R. & AVIÑA-ZUBIETA, J. A. 2015. Approaches for estimating minimal clinically important differences in systemic lupus erythematosus. *Arthritis Res Ther,* 17.

RAND. 2015. *36-Item Short Form Survey from the RAND Medical Outcomes Study* [Online].Available: http://www.rand.org/health/surveys_tools/mos/mos_core_36item.html [Accessed 02/09 2015].

RFPB. 2012. *Guidance Information for Applicants* [Online]. Available: http://www.ccf.nihr.ac.uk/RfPB/Documents/RfPB%20-%20Guidance%20for%20Applicants%20Competition%2018.pdf [Accessed 16th October 2012].

SANDVIK, L., ERIKSSEN, J. A. N., MOWINCKEL, P. & RØDLAND, E. A. 1996. A Method for Determining the Size of Internal Pilot Studies. *Statistics in Medicine,* 15**,** 1587-1590.

SCHOENFELD, D. 1980. Statistical Considerations For Pilot Studies. *International Journal of Radiation Oncology*Biology*Physics,* 6**,** 371-374.

SEELBINDER, B. M. 1953. On Stein's Two-stage Sampling Scheme. *The Annals of Mathematical Statistics,* 24**,** 640-649.

SIM, J. & LEWIS, M. 2011. The Size of a Pilot Study for a Clinical Trial Should be Calculated in Relation to Considerations of Precision and Efficiency. *Journal of Clinical Epidemiology*.

SIM, J. & LEWIS, M. 2012. The Size of a Pilot Study for a Clinical Trial should be Calculated in Relation to Considerations of Precision and Efficiency. *Journal of Clinical Epidemiology,* 65**,** 301-308.

SINGER, J. 1999. A Method for Determining the Size of Internal Pilot Studies by L. Sandvik, J. Erikssen, P. Mowinckel and E. A. Rødland, Statistics in Medicine, 15, 1587–1590 (1996). *Statistics in Medicine,* 18**,** 1151-1153.

SNEDECOR, G. W. & COCHRAN, W. G. 1989. *Statistical Methods,* Iowa, Iowa State University Press.

SPRANGE, K., MOUNTAIN, G., BRAZIER, J., COOK, S., CRAIG, C., HIND, D., WALTERS, S., WINDLE, G., WOODS, R., KEETHARUTH, A., CHATER, T. & HORNER, K. 2013. Lifestyle Matters for maintenance of health and wellbeing in people aged 65 years and over: study protocol for a randomised controlled trial. *Trials,* 14**,** 302.

STALLARD, N. 2011. Optimal Sample Sizes for Phase II Clinical Trials and Pilot Studies. *Statistics in Medicine*.

STALLARD, N., WHITEHEAD, J. & CLEALL, S. 2005. Decision-Making in a Phase II Clinical Trial: A New Approach Combining Bayesian and Frequentist Concepts. *Pharmaceutical Statistics,* 4**,** 119-128.

STEIN, C. 1945. A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance. *The Annals of Mathematical Statistics,* 16**,** 243-258.

STEVENS, J. W. 2009. *What is Bayesian Statistics?* [Online]. Available: http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/What_is_Bay_stats.pdf [Accessed 21st January 2016].

STRATFORD, P. W. 2010. The Added Value of Confidence Intervals. *Physical Therapy,* 90**,** 333-335.

SULLY, B. G. O., JULIOUS, S. A. & NICHOLL, J. 2013. A Reinvestigation of Recruitment to Randomised, Controlled, Multicenter Trials: A review of trial funded by two UK funding agencies. *Trials***,** 166.

SWINSCOW, T. D. V. & CAMPBELL, M. J. 2002. *Statistics at Square One*, BMJ Publishing Group.

TEAM MATH INC. 2011. Standard Normal Table. [Online] Available: http://www.normaltable.com/. [Accessed 30 October 2016].

TEARE, M. D., DIMAIRO, M., SHEPHARD, N., HAYMAN, A., WHITEHEAD, A. & WALTERS, S. J. 2014. Sample Size Requirements to Estimate Key Design Parameters from External Pilot Randomised Controlled Trials: A Simulation Study. *Trials,* 15.

TEIJLINGEN, E. R. V. & HUNDLEY, V. 2001. The Importance of Pilot Studies. *Social Research Update,* 35.

TEMPLE, R. 1999. Are Surrogate Markers Adequate to Assess Cardiovascular Disease Drugs? *The Journal of American Medical Association,* 282**,** 790-795.

THABANE, L., MA, J., CHU, R., CHENG, J., ISMAILA, A., RIOS, L. P., ROBSON, R., THABANE, M., GIANGREGORIO, L. & GOLDSMITH, C. H. 2010. A Tutorial on Pilot Studies: The What, Why and How. *BMC Medical Research Methodology,* 10.

TORGERSON, D. J. & TORGERSON, C. J. 2008. Designing Randomised Trials in Health, Education and the Social Sciences: An introduction. Basingstoke: Palgrave Macmillan.

VICKERS, A. J. 2003. Underpowering in Randomized Trials Reporting a Sample Size Calculation. *Journal of Clinical Epidemiology,* 56**,** 717-720.

WATSON, J. M. & TORGERSON, D. J. 2006. Increasing recruitment to randomised trials: a review of randomised controlled trials. *BMC Med Res Methodol,* 6**,** 34.

WHITEHEAD, A. L., SULLY, B. G. O. & CAMPBELL, M. J. 2014. Pilot and Feasibility Studies: Is there a difference from each other and from a randomised controlled trial? *Contemporary Clinical Trials,* 38**,** 130-133.

WHITEHEAD, A. L., JULIOUS, S. A., COOPER, C. L. & CAMPBELL, M. J. 2015. Estimating the Sample Size for a Pilot Randomised Trial to Minimise the Overall Trial Sample Size for the External Pilot and Main Trial for a Continuous Outcome Variable. *Statistical Methods in Medical Research.*

WHITEHEAD, J. 1997. *The Design and Analysis of Sequential Clinical Trials,* Chichester, John Wiley & Sons, Ltd.

WHO. 2015. *Health Topics: Health Technology* [Online]. Available: http://www.who.int/topics/technology_medical/en/ [Accessed 02/09 2015].

WITTES, J. 2002. Sample Size Calculations for Randomized Controlled Trials. *Epidemiologic Reviews,* 24**,** 39-53.

WITTES, J. & BRITTAIN, E. 1990. The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials. *Statistics in Medicine,* 9**,** 65-72.

WITTES, J., SCHABENBERGER, O., ZUCKER, D., BRITTAIN, E. & PROSCHAN, M. 1999. Internal Pilot Studies I: Type I Error Rate of the Naive t-test. *Statistics in Medicine,* 18**,** 3481-3491.

WRIGHT, A., HANNON, J., HEGEDUS, E. J. & KAVCHAK, A. E. 2012. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther,* 20**,** 160-6.

ZUCKER, D., WITTES, J. T., SCHABENBERGER, O. & BRITTAIN, E. 1999. Internal Pilot Studies II: Comparison of Various Procedures. *Statistics in Medicine,* 18**,** 3493-3509.

Appendix A – Statistical Tests

There are many statistical tests available for the analysis of data. Choosing the correct statistical test to be used for the analysis of data depends on the design or aim of the trial, the type of data, the number of covariates and the number of treatment groups.

The concentration in this thesis is on superiority trials with two independent treatment groups and a continuous endpoint, which will be assumed to be Normally distributed. The statistical tests, which relate to the sample size calculations discussed in Chapter 2 are presented here.

A.1 Z-Test

The Z-test allows the comparison of the treatment effect between two groups of Normally distributed data when no covariates will be taken into account. This test can be used in situations where the standard deviation for the population is known (Kirkwood and Sterne, 2003).

For the Z-test the test statistic, denoted as, Z is calculated from,

$$Z = \frac{\bar{x}_A - \bar{x}_B}{SE}, \tag{9.1}$$

where $\bar{x}_A$ is the sample mean from group A with variance s_A^2, $\bar{x}_B$ is the sample mean of group B with variance s_B^2 and the standard error is,

$$SE = \sqrt{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)}. \tag{9.2}$$

where n_A is the number of participants in group A and n_B is the number of participants in group B. This test statistic will be compared to a Standard Normal distribution table (a Normal distribution with a mean of 0 and standard deviation of 1), in order to find the p-value.

The confidence interval for the treatment effect when using a Z-test is given by the formula set out below:

$$(\bar{x}_A - \bar{x}_B) \pm Z_{(1-\alpha/2)}SE,$$

where $Z_{(1-\alpha/2)}$ is the standard Normal Z-score of (1 – α/2), α is usually set at 0.05 to represent a 95% confidence interval here, $Z_{(1-\alpha/2)}$ would equal 1.96.

A.2 Independent Samples T-Test

The independent samples t-test also allows the comparison of the treatment effect between two independent groups of Normally distributed data when no covariates will be taken into account. This test however is for situations where the standard deviation for the population is unknown (Kirkwood and Sterne, 2003). This is most likely to be the situation in most cases.

For the T-test the test statistic, denoted by t, is calculated from,

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE}, \tag{9.3}$$

where $\bar{x}_A$ is the mean response in group A and $\bar{x}_B$ is the mean response in group B. However, this time the standard error is calculated from the formula

$$SE = s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}, \tag{9.4}$$

where s is the standard deviation estimate based on pooling the data from the two groups using the following formula

$$s = \sqrt{\left[\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A + n_B - 2)}\right]}, \tag{9.5}$$

This is a weighted average of the two estimates of the standard deviation weighted towards the group with the larger sample size.

The test statistic in this case will be compared to a t-distribution dependent on the number of degrees of freedom for the test. The degrees of freedom are defined by $(n_1 + n_2 - 2)$.

The confidence interval for the treatment effect will be given by:

$$(\bar{x}_1 - \bar{x}_2) \pm \mathrm{t}_{(df,1-\alpha/2)}SE,$$

The assumptions of an independent samples t-test are as follows:

- The standard deviations of the distributions from which the two samples are drawn should be equal,
- The observations should be independent of each other,
- The data should be Normally distributed within group (Swinscow and Campbell, 2002).

Appendix B – Normal Distribution Table

The table on the following page contains the standard Normal distribution. That is a Normal distribution with a mean of zero and a standard deviation of 1. Z is a standard Normal random variable. The tabulated values represent the value of the cumulative Normal distribution at z and give P(z < Z) (Team Math Inc., 2011).

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Appendix C – The CACTUS Trial

After a stroke some patients develop aphasia, meaning that they have difficulty communicating with others. People can show improvements in their ability for many years after the development of aphasia and so it is thought that some self-managed computer software could offer long term support for sufferers of aphasia.

CACTUS is therefore a pilot trial looking at the feasibility and acceptability of the designed intervention and the intervention effect on the word-finding ability of the participant. The intervention was 5 months of the self-managed word finding therapy, which was to be tested against the control group that continued to receive the usual language stimulation activities.

A total of 28 people completed the study: 13 in the control group and 15 in the intervention group. The clinical outcome measure was the number of words the participants were able to name correctly.

The data from the trial suggests that a full randomised controlled trial would be possible and warranted as the treatment group showed a statistically significant improvement in the naming of words compared to the control group (Palmer et al., 2011).

Appendix D – Programming Code

D.1 Function to Calculate the Sample Size per arm according the Non-Central t-distribution Approach

```
# Input Parameters:
#   a = type I error
#   b= type II error
#   d = required difference
#   s = standard deviation from pilot trial
#   r = allocation ratio between groups
#   m = degrees of freedom from the pilot trial

iterative<-function(d,s,r,a,b,m){

 #Create Matrices
 mat<-matrix(data=NA,4,1,byrow=T)
 mat2<-matrix(data=NA,2,1001,byrow=T)

 #m is number of degrees of freedom from pilot study
 mat[1,1]<-m

 #second row is equal to starting value from normal approximation

 z=qnorm(1-(a/2))

 brac=qt((1-b),df=m,ncp=z)

 bracsq=(brac^2)

 top=(r+1)*(s^2)*bracsq

 bottom=r*(d^2)

 n1=top/bottom

 x=floor(n1)

 mat[2,1]=x

 if (x<2) {x<-2}

 #iterate n up from the approximation until n>=equation

 j=0

 for (i in seq(x,(x+100),by=0.1)){

  j=j+1

  mat2[1,j]=i

  degf=(i*(r+1))-2
```

```
  t=qt((1-(a/2)),df=degf)


  brac2=qt((1-b),df=m,ncp=t)

  brac2sq=(brac2^2)

  top2=(r+1)*(s^2)*brac2sq

  bottom2=r*(d^2)

  n2=top2/bottom2

  x2=(n2)

  mat2[2,j]=x2

  mat[3,1]=x2

  mat[4,1]=x2

  #Calculating the power way

  t2=qt((1-(a/2)),df=degf)


  squrttop=r*i*(d^2)
  squrtbottom=(r+1)*(s^2)
  squrt=sqrt(squrttop/squrtbottom)

  q=pt(squrt,df=m,ncp=t2)

  k=1-q

  if (k>=(1-b)) break

 }


 return(mat[4,1])

}
```

D.2 Example Code to find the Minimum Trial Sample Sizes Based on the NCT Approach

```
# Input Parameters:
#   a = type I error
#   b = type II error
#   d = required difference
#   s = standard deviation from pilot trial
#   r = allocation ratio between groups
#   i = degrees of freedom from the pilot trial

a<-0.05
b<-0.1
d<-0.5
s<-1
r<-1

mat<-matrix(data=NA,3,125,byrow=T)              #Set up matrix for the results

j<-0                                             #loop counter
for(i in seq(2,250,by=2)){                       #loop through the pilot trial degrees of freedom
  j<-j+1
  mat[1,j]<-i+2                                  #overall two-arm pilot trial sample size
  mat[2,j]<-2*ceiling(iterative(d,s,r,a,b,i))    #overall two-arm main trial sample size
  mat[3,j]<-mat[2,j]+mat[1,j]                    #overall two-arm total trial sample size
}


min<-min(mat[3,])                                #the minimum overall total sample size
pos<-which.min(mat[3,])                          #which position in the matrix is the minimum
main<-mat[2,pos]                  #the main trial sample size that leads to the overall minimum sample size
pilot<-mat[1,pos]                 #the pilot trial sample size that leads to the overall minimum sample size
```

D.3 Example Code to find the Minimum Trial Sample Sizes Based on the UCL Approach

```
# Input Parameters:
#   a = type I error
#   b= type II error
#   d = required difference
#   s = standard deviation from pilot trial
#   r = allocation ratio between groups
#   i = degrees of freedom from the pilot trial
#   X = confidence level for the UCL approach

a=0.05
b=0.1
d=0.5
s=1
r=1
X=0.8

mat<-matrix(data=NA,3,125,byrow=T)          #Set up matrix for the results

j=0                                         #loop counter
for(i in seq(2,250,by=2)){                  #loop through the pilot trial degrees of freedom
 j=j+1
 mat[1,j]<-i+2                              #overall two-arm pilot trial sample size
 UCL<-(i*(s^2))/(qchisq((1-X),df=i))
 brac<-(qnorm(1-(a/2))+qnorm(1-b))^2
 top<-(r+1)*(UCL)*brac
 bot<-r*(d^2)
 recalss<-ceiling(top/bot)          #sample size recalculation - one-arm sample size
 mat[2,j]<-recalss*2                #sample size recalculation - two-arm sample size
 mat[3,j]<-mat[2,j]+mat[1,j]        #overall two-arm total trial sample size

}


min=min(mat[3,])                    #the minimum overall total sample size
pos=which.min(mat[3,])              #which position in the matrix is the minimum
main=mat[2,pos]                     #the main trial sample size that leads to the overall minimum sample size
pilot=mat[1,pos]                    #the pilot trial sample size that leads to the overall minimum sample size
```

D.4 Example Code to find the Trial Sample Sizes Based on using a Proportional Pilot Trial for the NCT Approach

```
#Input Parameters

#d - treatment difference
#s - standard deviation
#r - allocation ratio between treatment groups
#a - type I error
#b - type II error

d=0.2
s=1
r=1
a=0.05
b=0.1

#Create Matrix

mat<-matrix(data=NA,4,16,byrow=T)

#Normal Sample Size Calculation

n=(2*(((r+1)*(s^2)*((qnorm(1-(a/2))+qnorm(1-b))^2))/(r*(d^2))))

mat[2,1]=n

#Iterative Step

for (i in 1:15){

 mat[1,i]<-i

 mat[3,i]=max(20,(0.03*mat[2,i]))

 mat[4,i]=mat[2,i]+mat[3,i]

 mat[2,(i+1)]=iterative(d,s,r,a,b,(mat[3,1])-2)

}

mat
```

D.5 Example Code to find the Trial Sample Sizes Based on using a Proportional Pilot Trial for the UCL Approach

```
#Input Parameters

#d - treatment difference
#s - standard deviation
#r - allocation ratio between treatment groups
#a - type I error
#b - type II error

d=0.05
s=1
r=1
a=0.05
b=0.1

#Create Matrix

mat<-matrix(data=NA,4,16,byrow=T)

#Normal Sample Size Calculation

n=(2*(((r+1)*(s^2)*((qnorm(1-(a/2))+qnorm(1-b))^2))/(r*(d^2))))

mat[2,1]=n

#Iterative Step

for (i in 1:15){

 mat[1,i]<-i

 mat[3,i]=max(20,(0.03*mat[2,i]))

 mat[4,i]=mat[2,i]+mat[3,i]

 ucl80=((mat[3,i]-2)*(s^2))/(qchisq(0.2,df=(mat[3,i]-2)))

 mat[2,(i+1)]=(2*(((r+1)*(ucl80)*((qnorm(1-(a/2))+qnorm(1-b))^2))/(r*(d^2))))


}
```

D.6 Example Code to find Minimum Overall Cost of Trial and Sample Sizes Required using NCT Approach

```
# Input Parameters:
#   a = type I error
#   b= type II error
#   d = required difference
#   s = standard deviation from pilot trial
#   r = allocation ratio between groups
#   R = relative cost of pilot versus main trial


d=0.2                #required difference
s=1                  #standard deviation
r=1                  #allocation ratio
a=0.05               #type I error rate
b=0.1                #type II error rate


R<-0.5               #relative cost
Y<-c(1:100)          #vector of degrees of freedom for the variance estimate


#Set up matrix for results
 mat2<-matrix(data=NA,4,100,byrow=T)
 g=0                                              #loop counter

 for(i in seq_along(Y)){
  g=g+1                                           #loop counter
  mat2[1,g]<-i+1                                  #pilot sample size per arm
  mat2[2,g]<-iterative(d,s,r,a,b,(2*i))           #main sample size per arm
  mat2[3,g]<-(2*mat2[1,g])+(2*mat2[2,g])          #overall sample size - two arms
  mat2[4,g]<-(R*2*(mat2[1,g]))+(2*mat2[2,g])      #function to be minimised

 }


 mini=min(mat2[4,])                 #the minimum of row 4
 pos=which.min(mat2[4,])            #postion of this minimum
 main=2*ceiling(mat2[2,pos])        #main sample size - 2 arms
 pilot=2*mat2[1,pos]                #pilot sample size - 2 arms
 overall=pilot+main                 #overall trial size - 2 arms


R                    #relative cost
pilot                #pilot trial sample size - 2 arms
main                 #main trial sample size - 2 arms
overall              #overall sample size - 2 arms
```

D.7 Example Code to find Minimum Overall Cost of Trial and Sample Sizes Required using UCL Approach

```
# Parameters:
#   a = type I error
#   b= type II error
#   d = required difference
#   s = standard deviation from pilot trial
#   r = allocation ratio between groups
#   R = relative cost of pilot versus main trial


d=0.5               #required difference
s=1                 #standard deviation
r=1                 #allocation ratio
a=0.05              #type I error rate
b=0.1               #type II error rate


R<-0.5              #relative cost
Y<-c(1:100)         #vector of degrees of freedom for the variance estimate


#Set up matrix for results
mat2<-matrix(data=NA,4,100,byrow=T)
g=0                                                                  #loop counter

for(i in seq_along(Y)){
 g=g+1                                                               #loop counter
 mat2[1,g]<-i+1                                                      #pilot sample size per arm
 ucl80=((2*mat2[1,g])*s)/(qchisq(0.2,df=(2*mat2[1,g])))              #80% UCL value
 n80<-((r+1)*(ucl80)*((qnorm(1-(a/2))+qnorm(1-b))^2))/(r*(d^2))      #Main trial sample size per arm
 mat2[2,g]<-2*n80                                                    #Main sample size - two arms
 mat2[3,g]<-(2*mat2[1,g])+mat2[2,g]                                  #Overall sample size - two arms
 mat2[4,g]<-(R*2*(mat2[1,g]))+mat2[2,g]                              #function to be minimised

}


mini=min(mat2[4,])          #The minimum of row 4
pos=which.min(mat2[4,])     #Postion of this minimum
main=ceiling(mat2[2,pos])   #Main sample size - 2 arms
pilot=2*mat2[1,pos]         #Pilot sample size - 2 arms
overall=pilot+main          #Overall trial size - 2 arms


R                   #Relative cost
pilot               #Pilot trial sample size - 2 arms
main                #Main trial sample size - 2 arms
overall             #Overall sample size - 2 arms
```

D.8 Example Code to Investigate the Effect of the Internal Pilot Trial Design on the Power of the Main Trial

```
d=0.8                       #Standardised effect size
nompower=0.9                #Original power
alpha=0.05                  #Type I error
recalpower=0.9              #Power at sample size recalculation
nomvar=1                    #Nominal variance estimate
truevar=1                   #True variance
r=1                         #Allocation ratio between groups

brac1=(qnorm(1-(alpha/2))+qnorm(nompower))^2
top1=(r+1)*(nomvar)*brac1
bot1=r*(d^2)
nomss=ceiling(top1/bot1)            #Original sample size calculation
msdf=(2*nomss)-2                    #Degrees of freedom for original sample size calculation

#set up matrix
mat<-matrix(data=NA, 4, 999, byrow=T)

z=0

#for loop for percentiles
for(i in seq(from=0.001,to=0.999,by=0.001)){

z=z+1

 #IPT sample size
 IPTss=10                   #One-arm sample size
 totaldf=(2*IPTss)-2        #Two-arm sample size

 #Estimate variance
 chi<-qchisq(i,totaldf)     #generate a quantile of the chisq distribution
 frac<-chi/totaldf          #calculate the estimate of the sample variance based on this quantile
 pervar<-frac*truevar

 #Re-estimate sample size
 brac2=(qnorm(1-(alpha/2))+qnorm(recalpower))^2
 top2=(r+1)*(pervar)*brac2
 bot2=r*(d^2)
 recalss=ceiling(top2/bot2)         #sample size recalculation - one-arm sample size
 totalss=recalss*2                  #sample size recalculation - two-arm sample size

 #Restricted procedure
 endss<-if (recalss>nomss) recalss else nomss

 #Calculating the power
 top3<-r*(endss)*(d^2)
 bot3<-(r+1)*(truevar)
 power<-pnorm(sqrt(top3/bot3)-1.96)         #power if not adjusted upwards


 ind<-if (recalss>nomss) 1 else 0           #indicator 1 if sample size goes up at interim

 #Results
 mat[1,z]<-i
```

```
  mat[2,z]<-power
  mat[3,z]<-endss
  mat[4,z]<-ind

}

results<-c(rowMeans(mat[2:4,],na.rm=TRUE,dims=1))        #average power, sample size and indicator value
results

sd(mat[3,])
sd(mat[2,])
```

D.9 Example Code to Simulate a Trial to Investigate the Effect of the Internal Pilot Trial Design on the Power of the Main Trial

```
d=0.8                #Standardised effect size
nompower=0.9         #Original power
alpha=0.05           #Type I error
recalpower=0.9       #Power at sample size recalculation
nomvar=1.5           #Nominal variance estimate
truevar=1            #True variance
r=1                  #Allocation ratio between groups

brac1=(qnorm(1-(alpha/2))+qnorm(nompower))^2
top1=(r+1)*(nomvar)*brac1
bot1=r*(d^2)
nomss=ceiling(top1/bot1)          #Original sample size calculation
msdf=(2*nomss)-2                  #Degrees of freedom for original ss calculation

#set up matrix
mat<-matrix(data=NA, 4, 100000, byrow=T)

#for loop for simulations
for(i in 1:100000){


 #IPT sample size
 IPTss=ceiling(0.75*nomss)        #One-arm sample size
 totalp=2*IPTss                   #Two-arm sample size

 #simulate pilot
 pilot0<-rnorm(IPTss,mean=0,sd=sqrt(truevar))   #control arm
 pilot1<-rnorm(IPTss,mean=d,sd=sqrt(truevar))   #treatment arm
 PS=c(pilot0,pilot1)                            #combine the data from both arms

 #Estimate blinded variance
 vari1<-var(PS)           #unadjusted estimate of the variance
 top<-totalp
 bot<-4*(totalp-1)
 frac<-top/bot
 bias<-frac*(d^2)         #estimate of the bias of the unadjusted variance estimate
 blindvar=vari1-bias      #blinded estimate of the variance

 #Re-estimate sample size
 brac2=(qnorm(1-(alpha/2))+qnorm(recalpower))^2
 top2=(r+1)*(blindvar)*brac2
 bot2=r*(d^2)
 recalss=ceiling(top2/bot2)       #sample size recalculation - one-arm sample size
 totalss=recalss*2                #sample size recalculation - two-arm sample size

 #Restricted procedure
 endss<-if (recalss>nomss) recalss else nomss

 #Calculating the power
 top3<-r*(endss)*(d^2)
 bot3<-(r+1)*(truevar)
 power<-pnorm(sqrt(top3/bot3)-1.96)          #power of the trial

 ind<-if (recalss>nomss) 1 else 0            #indicator 1 if sample size goes up at interim
```

```
  #Results
  mat[1,i]<-i
  mat[2,i]<-power
  mat[3,i]<-endss
  mat[4,i]<-ind

}
```

D.10 Example Code to Investigate the Effect of the Internal Pilot Trial Design on the Power of the Main Trial Assuming Variance Unknown at both the Sample Size Recalculation and in the Original Calculation

```
d=0.2                #Standardised effect size
alpha=0.05           #Type I error
recalpower=0.9       #Power at sample size recalculation
nompower=0.9         #Nominal Power
truevar=1            #True variance
r=1                  #Allocation ratio between groups


#EPT Sample Size
EPTss<-10
EPTdf<-(2*EPTss)-2

mat2<-matrix(data=NA, 3, 999, byrow=T)
for(j in seq(from=0.1, to=99.9,by=0.1)){

 #Estimate Initial Variance
 chi2<-qchisq((j/100),EPTdf)
 frac2<-chi2/EPTdf
 var<-frac2*truevar

 ss=pwr.t.test(d=d/var,sig.level=alpha,power=nompower,type="two.sample",alternative="two.sided")
 nomss=ceiling(ss$n)                    #Original sample size calculation
 msdf=(2*nomss)-2                       #Degrees of freedom for original ss calculation

 #set up matrix
 mat<-matrix(data=NA, 4, 999, byrow=T)

 #for loop for percentiles
 for(i in seq(from=0.1,to=99.9,by=0.1)){


  #IPT sample size
  IPTss=10                  #One-arm sample size
  IPTdf=(2*IPTss)-2         #Two-arm sample size

  #Estimate variance
  chi<-qchisq((i/100),IPTdf)            #generate a quantile of the chisq distribution
  frac<-chi/IPTdf                       #calculate the estimate of the sample variance based on this
  pervar<-frac*truevar

  #Re-estimate sample size
  ss1=pwr.t.test(d=d/pervar,sig.level=alpha,power=recalpower,type="two.sample",alternative="two.sided")
  recalss=ceiling(ss1$n)    #sample size recalculation - one-arm sample size
  totalss=recalss*2         #sample size recalculation - two-arm sample size

  #Restricted procedure
  endss<-if (recalss>nomss) recalss else nomss

  #Calculating the power
  power<-pwr.t.test(n=endss,d=d/truevar,sig.level=alpha,type="two.sample",alternative="two.sided")
```

```
#power if not adjusted upwards

   ind<-if (recalss>nomss) 1 else 0                  #indicator 1 if sample size goes up at interim

   #Results
   mat[1,i*10]<-i
   mat[2,i*10]<-power$power
   mat[3,i*10]<-endss
   mat[4,i*10]<-ind

  }

  results<-c(rowMeans(mat[2:4,],na.rm=TRUE,dims=1))       #average power, sample size and indicator value
  results

  mat2[1,(j*10)]<-EPTss
  mat2[2,(j*10)]<-results[[2]]
  mat2[3,(j*10)]<-results[[1]]

}

results2<-c(rowMeans(mat2,na.rm=TRUE,dims=1))
results2
```

Appendix E – Papers Contributed to During PhD

This section includes copies of papers contributed to during the course of the PhD.

WHITEHEAD, A. L., JULIOUS, S. A., COOPER, C. L. & CAMPBELL, M. J. 2015. Estimating the Sample Size for a Pilot Randomised Trial to Minimise the Overall Trial Sample Size for the External Pilot and Main Trial for a Continuous Outcome Variable. *Statistical Methods in Medical Research.* – This paper is based on the work presented in Chapter 4 of this thesis. I carried out the work and drafted the paper.

WHITEHEAD, A. L., SULLY, B. G. O. & CAMPBELL, M. J. 2014. Pilot and Feasibility Studies: Is there a difference from each other and from a randomised controlled trial? *Contemporary Clinical Trials,* 38**,** 130-133. – I contributed to the work presented in this paper, largely based on the literature review I carried out for this thesis and helped to draft the manuscript.

TEARE, M. D., DIMAIRO, M., SHEPHARD, N., HAYMAN, A., WHITEHEAD, A. & WALTERS, S. J. 2014. Sample Size Requirements to Estimate Key Design Parameters from External Pilot Randomised Controlled Trials: A Simulation Study. *Trials,* 15. – I was involved in the work for this paper and reviewed the drafted manuscript.

LEE, E. C., WHITEHEAD, A. L., JACQUES, R. M. & JULIOUS, S. A. 2014. The Statistical Interpretation of Pilot Trials: Should significance thresholds be reconsidered? *BMC Medical Research Methodology,* 14**,** 41-41. – I contributed to the work presented in this paper and helped to draft the manuscript.

BILLINGHAM, S. A. M., WHITEHEAD, A. L. & JULIOUS, S. A. 2013. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Med Res Methodol,* 13**,** 104. – I co-supervised a medical student while they completed this project, I helped with the analysis and drafting the paper.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–17

DOI: 10.1177/0962280215588241
smm.sagepub.com
SAGE

Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable

Amy L Whitehead,[1] Steven A Julious,[1] Cindy L Cooper[2] and Michael J Campbell[1]

Abstract

Sample size justification is an important consideration when planning a clinical trial, not only for the main trial but also for any preliminary pilot trial. When the outcome is a continuous variable, the sample size calculation requires an accurate estimate of the standard deviation of the outcome measure. A pilot trial can be used to get an estimate of the standard deviation, which could then be used to anticipate what may be observed in the main trial. However, an important consideration is that pilot trials often estimate the standard deviation parameter imprecisely. This paper looks at how we can choose an external pilot trial sample size in order to minimise the sample size of the overall clinical trial programme, that is, the pilot and the main trial together. We produce a method of calculating the optimal solution to the required pilot trial sample size when the standardised effect size for the main trial is known. However, as it may not be possible to know the standardised effect size to be used prior to the pilot trial, approximate rules are also presented. For a main trial designed with 90% power and two-sided 5% significance, we recommend pilot trial sample sizes per treatment arm of 75, 25, 15 and 10 for standardised effect sizes that are extra small ($\leq$0.1), small (0.2), medium (0.5) or large (0.8), respectively.

Keywords

Pilot trial, RCT, sample size, power, continuous outcome

[1]Medical Statistics Group, Design, Trials and Statistics Group, School of Health and Related Research, University of Sheffield, Sheffield, UK
[2]Clinical Trials Research Unit, Design, Trials and Statistics Group, School of Health and Related Research, University of Sheffield, Sheffield, UK

Corresponding author:
Amy L Whitehead, Medical Statistics Group, Design, Trials and Statistics Group, School of Health and Related Research, University of Sheffield, Sheffield, UK.
Email: a.whitehead@sheffield.ac.uk

1 Introduction

Sample size is an important consideration when a clinical trial is planned, not only for the main trial but also for any preliminary pilot trial. A sample size calculation is used to determine the minimum number of participants needed in a clinical trial in order to be able to answer the research question under investigation.[1] Recruiting too few participants in a main trial means that the probability of finding a clinically relevant difference statistically significant is low and as a consequence, the chance of inconclusive results is high.[2,3] Conversely, if the sample size is too large, resources may be wasted, more patients than necessary could be given a treatment which will later be proven to be inferior; or an effective treatment may be delayed from being released on to the market.[4]

For the purpose of this work, we are defining a pilot randomised trial as a trial, which mimics the design of the main trial but is not designed with the aim to prove the superiority of one treatment over another[5] but rather to try out aspects of the proposed main trial. As pilot trials do not have the same objectives as a main trial, setting the sample size in the same way – using formal power considerations – is usually not necessary. However, it is still necessary to provide a sample size justification even when the reasons for choosing a particular size are pragmatic.

The focus of this paper will be deriving pilot trial sample sizes based on a primary aim of the pilot being to estimate the standard deviation to be used for the main trial sample size calculation. We will describe a method for estimating the sample size for a pilot trial, which achieves the objective of minimising the recruitment of patients across the pilot and the main trial overall. The emphasis in this paper is on two armed superiority trials; however, the results are easily generalisable to trials with other designs. Furthermore, we will concentrate on external pilot trials where the assumption, however, is that there are no changes between it and the main trial, so that the standard deviation of the outcome measurement is unaffected. We are also not considering the situation of an internal pilot trial where the data are combined from the pilot trial and the main trial for the final analysis.

2 Standard methods

For a continuous normally distributed outcome, in a superiority trial, the sample size per treatment arm, n, to ensure adequate power $(1-\beta)$ where β is the Type II error rate whilst controlling the Type I error rate, α, for a specified/required treatment difference, d, and standard deviation, σ, is given by

$$\frac{n = (r+1)(z_{1-\beta} + z_{1-\alpha/2})^2\sigma^2}{rd^2} \tag{1}$$

where r is the allocation ratio of participants between the two treatment arms, experimental to control.[6]

Subjective clinical expertise can be used to specify the required treatment difference and there are agreed values used for the Type I and II error levels. However, a difficulty arises when trying to quantify the standard deviation.[7] Estimating the standard deviation at an inappropriate level can have a serious effect on the power of the study.[8] If the anticipated standard deviation is estimated to be too high, the trial will contain more participants than necessary. If the anticipated value is estimated to be too low, the trial will not contain enough participants to find the required effect, leading to the problems outlined in Section 1.

One of the methods investigators might use to try to get an accurate prediction of the true standard deviation (or variance) of the outcome measure is to conduct an external pilot trial prior to the main trial. Pilot trials are often small; therefore, they tend to imprecisely predict the true variance. The anticipated distribution of the pilot variance is a chi-squared distribution.[9] As a

consequence, the accuracy of the variance prediction will depend on the pilot sample size and, hence, the degrees of freedom for the variance. Estimating the main trial sample size from equation (1) can result in a loss of power when the variance is imprecisely estimated. Using previous trial results to estimate the variance introduces a type of imprecision that should be allowed for when estimating the sample size for the main trial.[9]

2.1 Adjusting the standard deviation estimate from a pilot trial

Two different methods have been developed to try to deal with the issue of imprecise variance estimates. The first was proposed by Browne[10] and will be referred to as the upper confidence limit (UCL) approach and the second by Julious and Owen[9] which will be referred to as the non-central t-distribution (NCT) approach. In both methods, the sample size is inflated to allow for the imprecision involved when estimating the variance of an outcome measure from a pilot trial.

2.1.1 UCL approach

The UCL approach uses an $100X\%$ UCL for the estimated value of the variance from the pilot trial to plan the main trial. Browne[10] contended that this provides a sample size sufficient to achieve the required power in at least $100X\%$ of such trials. Browne recommends an 80% upper confidence level. However, Sim and Lewis,[11] whose results will be discussed later in the paper, set X at 0.95 or the 95% level.

In order to implement the UCL approach, a variance estimate from the pilot data is obtained and the one-sided $X\%$ UCL for this variance, s^2_{UCL}, is calculated. A one-sided $100X\%$ UCL for the variance can be calculated from

$$s^2_{UCL} = \left[\frac{k}{\chi^2_{1-X,k}}\right] s^2 \tag{2}$$

where s^2 is the pooled variance from the pilot trial with k degrees of freedom for the variance estimate, and $\chi^2_{1-X,k}$ denotes the $1-X$ percentile of the chi-squared distribution with k degrees of freedom.[12] As k increases, the confidence interval for a variance estimate becomes narrower.

Note for a two arm parallel group pilot trial with equal allocation to treatments, k would usually be $k = 2m - 2$, where m is the sample size per arm in the pilot trial from which the variance is being estimated.

This UCL would, therefore, be used as the variance estimate in the traditional sample size equation given earlier in equation (1). Therefore, the sample size per treatment arm for the main trial, n_M, would be given by

$$n_M = \frac{(r+1)\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2 s^2_{UCL}}{rd^2}. \tag{3}$$

If we investigate how much larger the sample size estimate is from this approach compared to the standard approach, by dividing equation (3) by equation (1) with s^2 used as an estimate of σ^2, we find that the UCL approach sample size is larger by a factor of $[\frac{k}{\chi^2_{1-X,k}}]$. Therefore, the factor by which the UCL approach sample size is greater than the standard approach depends only upon the pilot trial sample size and the value of X. It is possible, therefore, to calculate inflation factors, which can be used to multiply by the sample size from a standard calculation to give the sample size for the

Table 1. Inflation factors for the sample size calculation using the UCL approach.

Pilot trial sample size	80% upper confidence limit	95% upper confidence limit
20	1.400	1.917
24	1.349	1.783
30	1.297	1.654
40	1.244	1.527
50	1.211	1.450
70	1.172	1.359
100	1.139	1.287
200	1.093	1.190

UCL approach for a set value of total pilot trial sample size and X; these can be seen in Table 1. The pilot trial sample sizes used here are total sample sizes across treatment arms – assuming for the purpose of this paper, the trial is a two armed trial.

2.1.2 NCT approach

Julious and Owen[9] suggest an alternative method for the calculation of sample size accounting for the fact that we are using a sample estimate of the variance rather than the population variance in the calculation. The sample size inflation is dependent on the number of degrees of freedom on which the estimate of the variance is based, k; therefore, the sample size per treatment arm for the main trial, n_M, would be given by

$$n_M \geq \frac{(r+1)\left[t^{-1}\left(1-\beta, k, t^{-1}(1-\alpha/2, n_M(r+1)-2, 0)\right)\right]^2 s^2}{rd^2} \tag{4}$$

where $t^{-1}(., k, a)$ is the inverse function of the cumulative distribution function of a NCT with a non-centrality parameter, a, on k degrees of freedom. The non-centrality parameter in this case is $t^{-1}(1-\alpha/2,\ n_M(r+1)-2,\ 0)$ which is the inverse function of the cumulative distribution function of a central t-distribution with $n_M(r+1)-2$ degrees of freedom (as a = 0). Here k is the degrees of freedom for the variance estimate s^2. If the estimate of the variance is based on only a few degrees of freedom, the sample size will be increased. Consequently, as the number of degrees of freedom for the estimate of the variance increases, the impact of this method on the sample size diminishes. As can be seen in the paper by Julious and Owen,[9] it is also possible to calculate inflation factors for the NCT approach. The inflation factor represents how much larger the NCT approach sample size would be compared to the standard sample size calculation. Table 2 shows the inflation factors for this approach for total pilot trial sample sizes.

The UCL approach inflation depends only on the pilot trial sample size and the chosen level of X, whereas the NCT inflation factor depends on the pilot trial sample size and the Type I and Type II error rates. We can see from Tables 1 and 2 that the inflation factors for the UCL approach when X is 80 or 95% are much higher than the inflation factors for the NCT approach. Table 3 demonstrates which value of X in the UCL approach would make the inflation factor equal to that of the NCT method, as well as the resulting inflation factor, the sample sizes presented are total pilot trial sample sizes.

Table 2. Inflation factors for the sample size calculation for the NCT approach when the Type I error is 5%.

	Power	
Pilot trial sample size	90%	80%
20	1.156	1.099
24	1.125	1.080
30	1.097	1.062
40	1.071	1.045
50	1.055	1.036
70	1.039	1.025
100	1.027	1.017
200	1.013	1.009

Table 3. Inflation factors and levels of X for the UCL approach that give the same sample size as the NCT approach.

	Power			
	90%		80%	
Pilot trial sample size	X	Inflation factor	X	Inflation factor
20	0.622	1.156	0.566	1.099
24	0.611	1.125	0.560	1.080
30	0.599	1.097	0.553	1.062
40	0.586	1.071	0.546	1.045
50	0.577	1.056	0.541	1.036
70	0.565	1.039	0.534	1.025
100	0.554	1.027	0.529	1.017
200	0.538	1.013	0.520	1.008

It can be seen that as the pilot sample size increases the value for X in the UCL approach, which would lead to the same sample size as the NCT approach tends towards 0.5 and the inflation factor tends towards 1.

2.2 Pilot trial sample sizes

So far, we have highlighted how to estimate the sample size for a main trial based on the estimates of variance from a pilot trial. The question now being considered is how to estimate the sample size for the pilot trial in the situation where the variance estimate from the pilot trial is being used to design a main trial.

As highlighted previously, in a pilot trial the objective is not to prove superiority of the treatment but to test trial procedures and processes and to get estimates of parameters for the main trial sample

size calculation.[13–15] Therefore, the sample size formulae which are used for main treatment assessments are not usually applicable to pilot trials. The Consolidated Standards of Reporting Trials Group and bodies such as The National Institute for Health Research and The National Research Ethics Service state that not all studies necessarily need a power-based sample size calculation but they do all need a sample size justification. Therefore, since the purpose of the pilot is not to give a formal assessment of efficacy, then the sample size provided by the conventional calculations may be higher than necessary.[13]

2.2.1 Rules of thumb

When estimating the sample size for the pilot trial, the simplest methods to apply are sample size rules of thumb. Browne[10] cites a general flat rule to 'use at least 30 subjects or greater to estimate a parameter', whereas Julious[16] suggests a minimum sample size of 12 subjects per treatment arm. Teare et al.[17] recommend a pilot trial sample size of 70 in order to reduce the imprecision around the estimate of the standard deviation. All of these rules have limitations, however, as they are applied regardless of the size of the main trial being designed. The cost of the simplicity of this flat approach, is a larger overall sample size when the main trial is large or small, as demonstrated in Section 4.

2.2.2 Minimising the sample size across studies approach

If one of the adjustment methods described in the previous section to account for imprecision in the variance estimation is applied to calculate the main trial sample size, it would mean that the pilot trial sample size would affect the sample size of the main trial. That is, the methods depend on the degrees of freedom around the variance estimate and hence the pilot sample size.

There is a trade-off, therefore, between having a small pilot study and a larger main trial or a larger pilot study and a smaller main trial. This is because the larger the pilot the more precisely estimated the variance will be and, hence, the smaller the inflation factor applied to the main study sample size calculation. However, eventually the pilot sample size will get too large, and the number included in the pilot trial will outweigh the reduction in the main trial sample size. Therefore, it may be appropriate to consider the implications of this relationship when choosing the sample size of the pilot trial.

The method of setting the pilot trial sample size in order to minimise the overall sample size of the pilot and the main trial together was described by Kieser and Wassmer.[12] They applied the 80% UCL approach to the sample size calculation and found that a pilot trial sample size between 20 and 40 would minimise the overall sample size for a main study sample size of 80–250 corresponding to standardised effect sizes of 0.4 and 0.7 (for 90% power based on a standard sample size calculation). Sim and Lewis[11] also applied the UCL approach in their work but with a 95% UCL. They found that a pilot trial of $n \geq 55$ would minimise the overall sample size for small to medium standardised effect sizes (0.2–0.6). The impact of Sim and Lewis' use of a 95% UCL is that it has the effect of increasing their estimate of the required sample size compared to Kieser and Wassmer, for both the pilot and the main trial.

The current methods for setting pilot trial sample sizes are based on a set of rules, which we will call flat rules of thumb, these are given in Table 4. These pilot sample sizes are fixed no matter how large the subsequent main trial will be.

Please note that the sample sizes presented in Tables 1 to 4 and in Figures 2 to 4 are the total sample size required for a two arm trial. This has been done to allow for comparisons to be made between the flat rules of thumb: as some rules are based on the numbers of participants required per arm and some are based on the total number of participants required – for example, Sim and Lewis[11]

Table 4. The current flat rules of thumb for overall pilot trial sample size of a two armed trial.

Author	Recommended pilot trial sample size
Julious[16]	24
Kieser and Wassmer[12]	20–40
Browne[10]	30
Sim and Lewis[11]	$\geq$55
Teare et al.[17]	70

recommend 55 or more patients in total. The results presented in Figure 1 and Tables 5, 6 and 8 are per treatment arm. This allows for generalisability to trials with two or more treatment arms.

2.3 Summary of standard methods

Setting the pilot trial sample size in order to minimise the total sample size of the pilot and the main trial together could be argued to be the most appropriate method of sample size calculation as it recognises that the pilot trial is part of a larger clinical development programme, rather than a stand-alone study. Other methods fail to recognise this point and aim to minimise both the pilot and the main trials separately which could lead to the suboptimal sample size overall.

3 Proposed methods of optimising the sample size across studies

Using standardised differences ($\delta = d/s$) and pilot trial sample sizes per treatment group of 1 and upwards, we can calculate the required main trial sample sizes based on all combinations of these variables using the NCT approach through equation (4). As n_M appears on both sides of equation (4), it can be solved iteratively. To calculate a starting point for the iterations we can use,

$$n_{START} = \frac{(r+1)s^2\left[t^{-1}\left(1-\beta, k, z_{1-\alpha/2}\right)\right]^2}{rd^2} \tag{5}$$

which gives a direct estimate of the sample size without iteration. Once the required main trial sample size per arm, n_M, has been found, it is then added to the specified pilot trial sample size per arm, m, $m = (k+2)/2$ for a two armed design, to find the overall study sample size per arm (N_O) if this design is to be used.

$$N_O = m + n_M \tag{6}$$

For each value of δ, the pilot trial sample size per arm, m_{OPT}, which minimises the size of the overall study, N_O, can be found; this is referred to as the optimal pilot trial sample size. Therefore, if the δ to be used in the main trial is known, it is possible to calculate exactly the optimal pilot trial sample size in order to minimise the overall trial sample size. This process is depicted in Figure 1.

However, the exact δ to be used in the main trial may not be known at this early stage. Therefore, pilot trial sample size rules of thumb have been calculated based on the small, medium or large standardised effect sizes as set out by Cohen.[18]

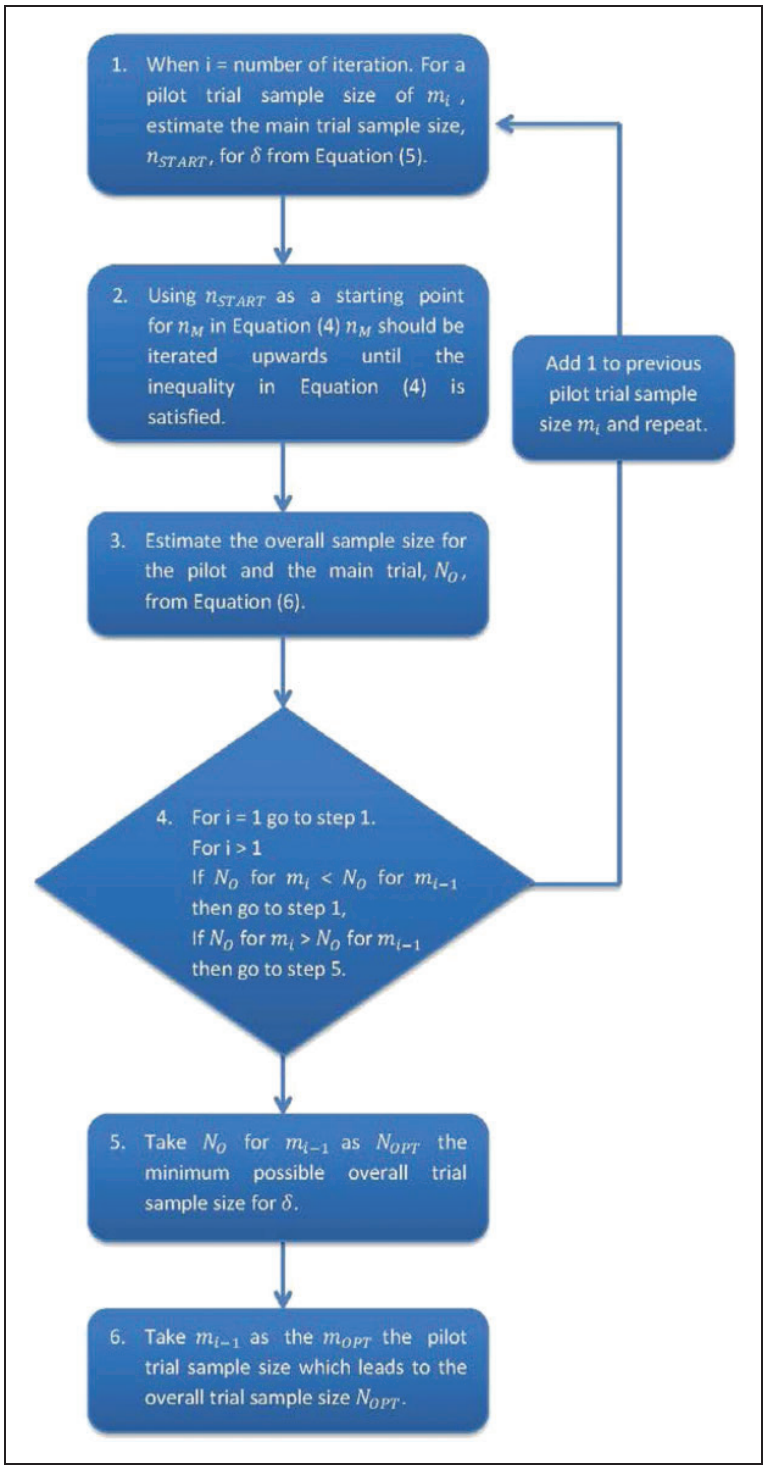


Figure 1. Process for calculating the optimal pilot trial sample size.

Table 5. Theoretical optimal values of pilot trial, main trial and overall trial sample size per treatment arm for each inflation method.

Standardised difference	Inflation method								
	80% upper confidence limit			95% upper confidence limit			Non-central t-distribution		
	Pilot	Main	Overall	Pilot	Main	Overall	Pilot	Main	Overall
80% powered main trial									
0.05	210	6671	6881	331	6892	7223	74	6353	6427
0.10	88	1728	1816	139	1817	1956	38	1607	1645
0.20	39	457	496	61	493	554	20	412	432
0.25	30	300	330	47	326	373	16	267	283
0.30	24	213	237	38	234	272	14	188	202
0.40	18	125	143	28	139	167	11	108	119
0.50	14	83	97	22	94	116	9	71	80
0.60	12	60	72	18	69	87	8	51	59
0.70	10	45	55	16	53	69	7	38	45
0.75	10	40	50	15	47	62	7	33	40
0.80	9	36	45	14	42	56	6	30	36
0.90	8	29	37	13	35	48	6	24	30
1.00	7	25	32	11	29	40	5	20	25
90% powered main trial									
0.05	253	8880	9133	398	9149	9547	106	8511	8617
0.10	106	2292	2398	167	2400	2567	54	2154	2208
0.20	46	603	649	72	647	719	28	552	580
0.25	35	394	429	56	427	483	23	358	381
0.30	29	279	308	45	305	350	19	252	271
0.40	21	163	184	33	181	214	15	145	160
0.50	16	108	124	26	122	148	12	95	107
0.60	14	78	92	21	89	110	11	68	79
0.70	12	59	71	18	68	86	9	51	60
0.75	11	52	63	17	60	77	9	45	54
0.80	10	46	56	16	54	70	8	40	48
0.90	9	38	47	14	44	58	8	32	40
1.00	8	32	40	13	37	50	7	27	34

4 Results

4.1 Optimal sample sizes

In order to calculate the minimum possible overall sample size for each standardised difference and adjustment method, the method presented in Figure 1 was used. The total sample size required for a two armed main trial for standardised differences of 0.2, 0.5 and 0.8 can be seen in Figures 2 to 4, these were calculated based on a power of 90%, Type I error rate of 5% and an allocation ratio, r, of 1.

It can be seen from Figures 2 to 4 that it is possible to solve the function and find the pilot trial sample size, which minimises the overall trial sample size. Table 2 shows the optimal pilot sample size, the required main trial sample size for the pilot trial and then the resulting overall trial sample

Table 6. Theoretical optimal values of pilot trial, main trial and overall trial sample size per treatment arm for each inflation method with a floor on the lower limit of pilot trial sample size at 10 per arm.

	Inflation method								
	80% upper confidence limit			95% upper confidence limit			Non-central t-distribution		
Standardised difference	Pilot	Main	Overall	Pilot	Main	Overall	Pilot	Main	Overall
80% powered main trial									
0.05	210	6671	6881	331	6892	7223	74	6353	6427
0.10	88	1728	1816	139	1817	1956	38	1607	1645
0.20	39	457	496	61	493	554	20	412	432
0.25	30	300	330	47	326	373	16	267	283
0.30	24	213	237	38	234	272	14	188	202
0.40	18	125	143	28	139	167	11	108	119
0.50	14	83	97	22	94	116	10	70	80
0.60	12	60	72	18	69	87	10	49	59
0.70	10	45	55	16	53	69	10	36	46
0.75	10	40	50	15	47	62	10	31	41
0.80	10	35	45	14	42	56	10	28	38
0.90	10	28	38	13	35	48	10	22	32
1.00	10	22	32	11	29	40	10	18	28
90% powered main trial									
0.05	253	8880	9133	398	9149	9547	106	8511	8617
0.10	106	2292	2398	167	2400	2567	54	2154	2208
0.20	46	603	649	72	647	719	28	552	580
0.25	35	394	429	56	427	483	23	358	381
0.30	29	279	308	45	305	350	19	252	271
0.40	21	163	184	33	181	214	15	145	160
0.50	16	108	124	26	122	148	12	95	107
0.60	14	78	92	21	89	110	11	68	79
0.70	12	59	71	18	68	86	10	50	60
0.75	11	52	63	17	60	77	10	44	54
0.80	10	46	56	16	54	70	10	39	49
0.90	10	37	47	14	44	58	10	31	41
1.00	10	30	40	13	37	50	10	25	35

size per treatment group for all adjustment methods based on a main trial power of 80%. Table 2 shows the same results but for a main trial power of 90%. The sample sizes presented in the tables are per treatment group.

The straight line on the graphs depicts a standard sample size calculation with no adjustment method applied (based on equation (1)). The points on the line show the resulting overall sample size if the rules of thumb of 24, 30 or 70 were used with no adjustment applied, the population variance is assumed to be known. The bottom dashed curve represents the NCT method as proposed by Julious and Owen.[9] The points on the line show the resulting overall trial sample size if the rules of thumb of 24 or 30 subjects were used for the pilot trial. The middle curve is the UCL method with an 80% UCL for the variance. The points represent the rules of thumb of 20 and 40 as set out by Kieser and

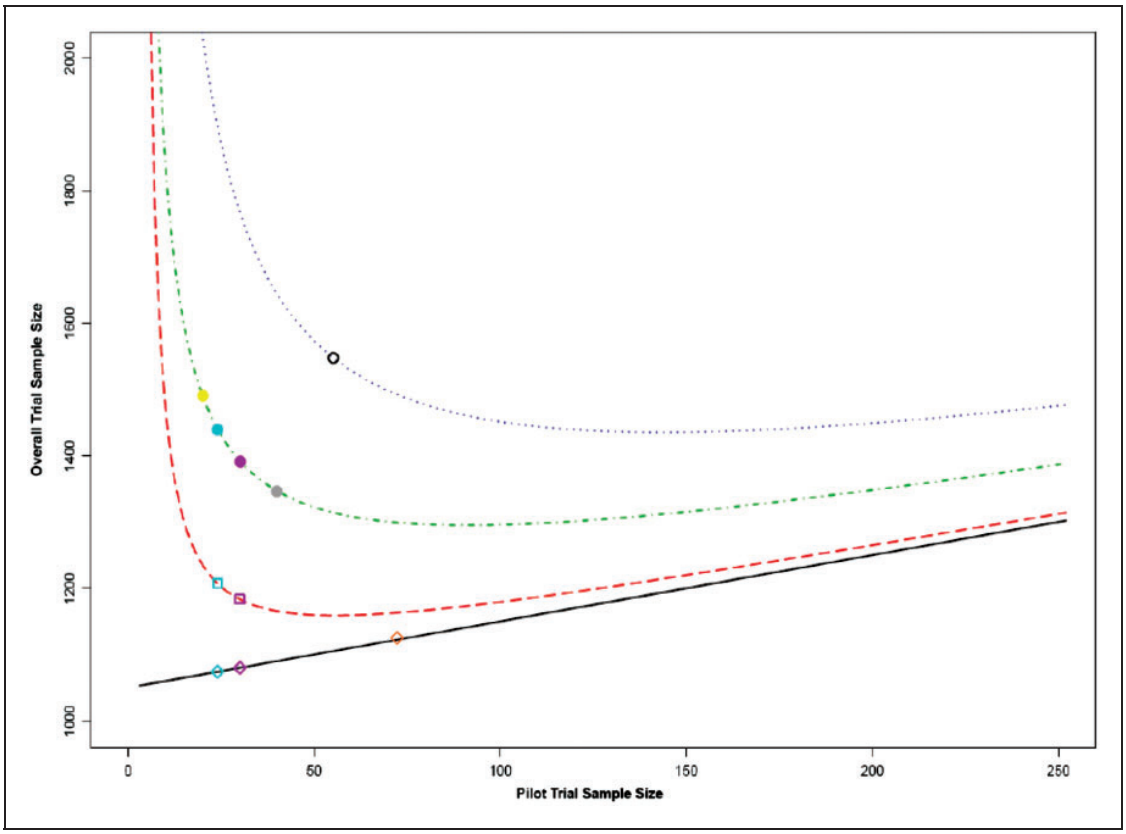


Figure 2. Comparing overall total trial sample sizes for each adjustment method over varying pilot trial sample size for a standardised difference of 0.2.

*Lines from bottom to top: Line 1, Standard sample size calculation with no adjustment method applied (points represent pilot trial sample sizes of 24, 30 and 70); Line 2, Main trial sample size calculation based on the NCT approach (points represent pilot trial sample sizes of 24 and 30); Line 3, Main trial sample size calculation based on the 80% UCL approach (points represent pilot trial sample sizes of 20, 24, 30 and 40) and Line 4, Main trial sample size calculation based on the 95% UCL approach (point represents pilot trial sample size of 55).

Wassmer[12] as well as the 24 and 30 rules. The top dotted curve is the UCL method with a 95% UCL for the variance. The point for a pilot trial sample size of 55 has been added here, as this was the sample size recommended by Sim and Lewis[11] to minimise the overall trial sample size. The overall trial sample sizes on the graphs are the total for a two armed trial.

The graphs in Figures 2 to 4 can be used to compare the effects of using the rules of thumb described in Table 2 to the theoretical optimal solution. For a medium standardised effect size (e.g., 0.5), the suggested rules of thumb are very close to the optimal pilot sample size. However, when the standardised effect size moves away from this value, the rules of thumb are less useful. For small standardised effect sizes (e.g., 0.2), the rules of thumb underestimate the required size of the pilot trial. For large standardised effect sizes (e.g., 0.8), the rules of thumb overestimate the number of participants required for the pilot trial. This indicates that the larger the main trial the larger the pilot trial should be in order to minimise the overall sample size; therefore; one fixed flat pilot trial sample size will not be suitable for all studies.

In relation to overall trial sample size, overestimating the pilot sample size is not as costly as underestimation in terms of over recruitment of participants as shown in Figures 2 to 4, given that

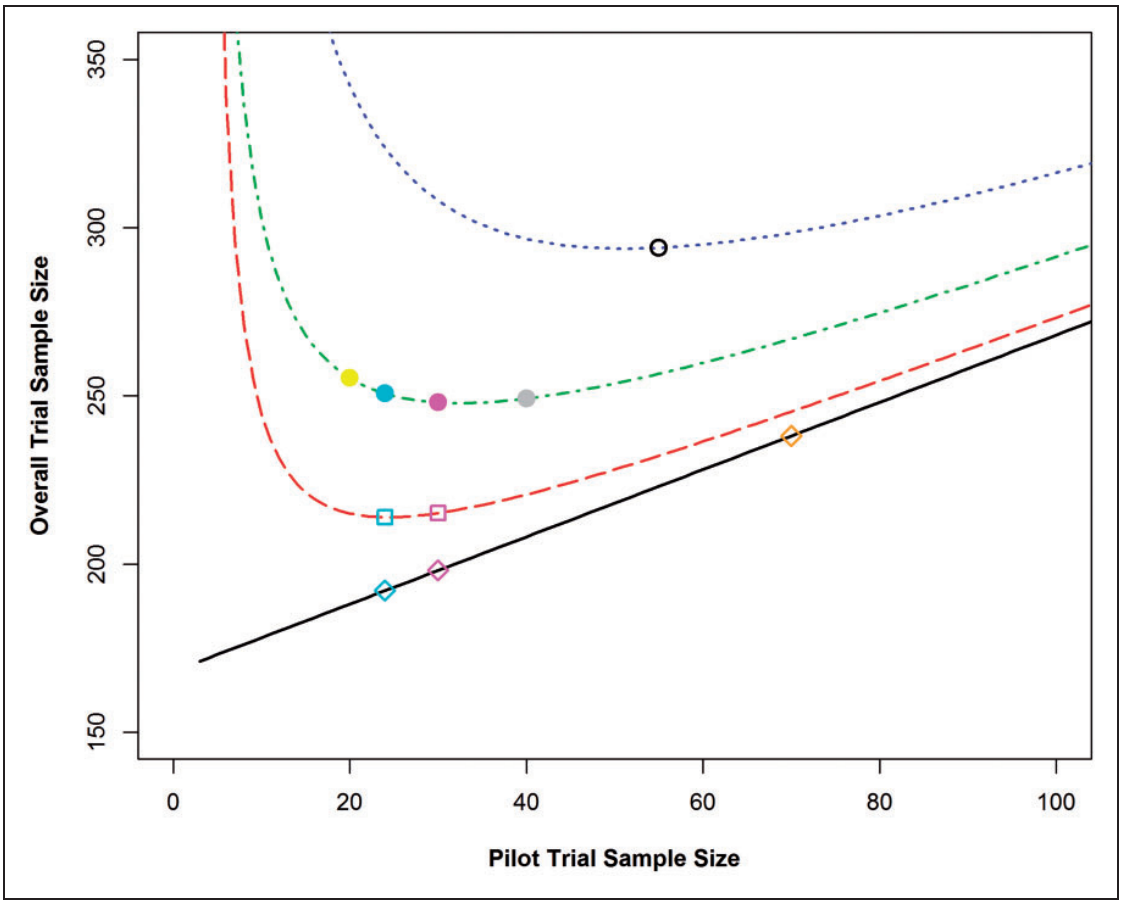


Figure 3. Comparing overall trial sample sizes for each adjustment method for varying pilot trial sample sizes for a standardised difference of 0.5.

*Lines from bottom to top: Line 1, Standard sample size calculation with no adjustment method applied (points represent pilot trial sample sizes of 24, 30 and 70); Line 2, Main trial sample size calculation based on the NCT approach (points represent pilot trial sample sizes of 24 and 30); Line 3, Main trial sample size calculation based on the 80% UCL approach (points represent pilot trial sample sizes of 20, 24, 30 and 40) and Line 4, Main trial sample size calculation based on the 95% UCL approach (point represents pilot trial sample size of 55).

the slope on the right hand side is flatter than on the left. The gradient of the slope on the right hand side of the minimum value is less than the gradient of the slope to the left side of the minimum; therefore, for the same change in pilot trial sample size – over estimation compared to underestimation – the change in overall trial sample size will be comparatively less. It can be seen that the NCT approach produces consistently lower overall trial sample sizes than any of the UCL methods.

It should be noted that for large values of standardised effect size, the suggested pilot trial sample size falls to a level, which may be considered too low to achieve the objectives of a pilot trial. This is because pilot trials are not only used to estimate the standard deviation of the outcome measure but also to assess objectives such as testing the feasibility of trial processes or predicting the likely dropout rate. We must consider these other objectives as well as more practical considerations. For the rest of this paper, a floor will be placed on the minimum pilot trial sample size per arm of 10 participants; this allows some investigation of these other objectives and is in line with the minimum sample size for an internal pilot trial sample size as recommended by Birkett and Day.[19]

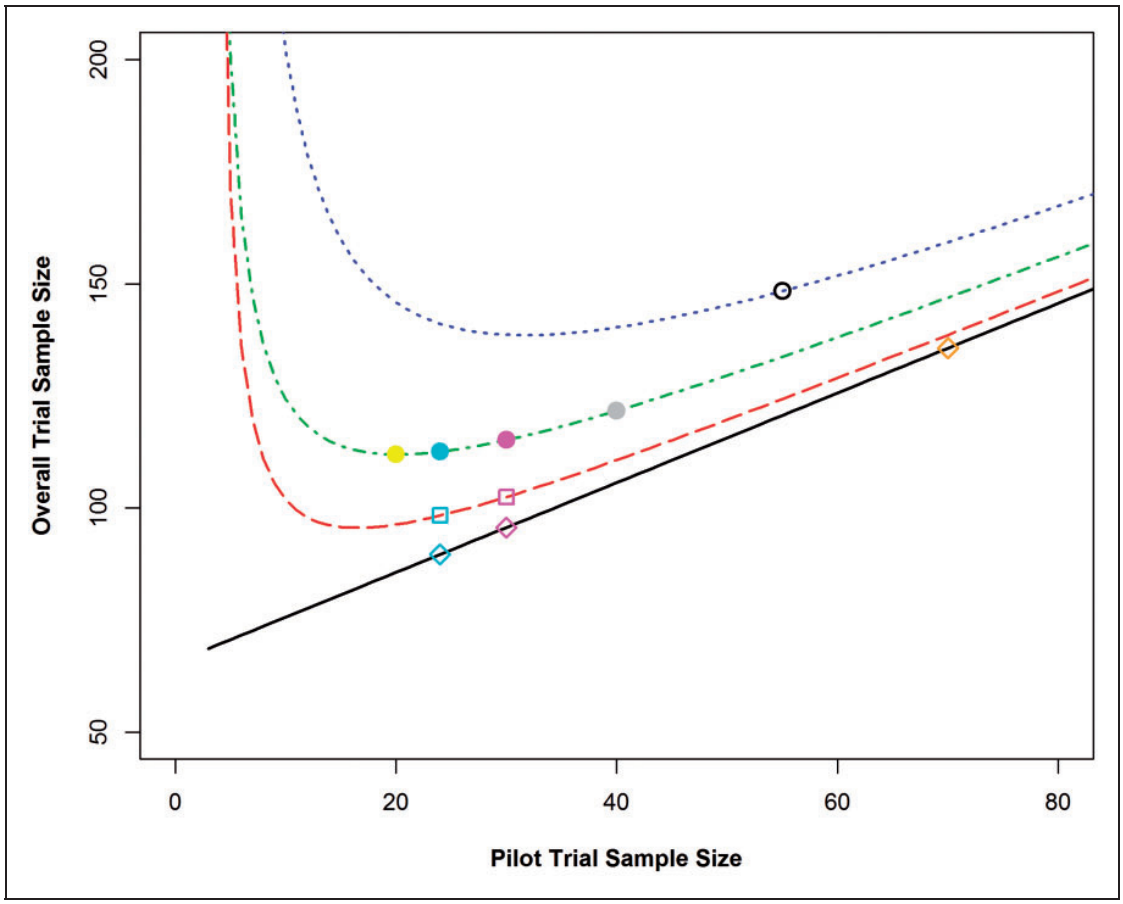


Figure 4. Comparing overall trial sample sizes for each adjustment method for varying pilot trial sample sizes for a standardised difference of 0.8.

*Lines from bottom to top: Line 1, Standard sample size calculation with no adjustment method applied (points represent pilot trial sample sizes of 24, 30 and 70); Line 2, Main trial sample size calculation based on the NCT approach (points represent pilot trial sample sizes of 24 and 30); Line 3, Main trial sample size calculation based on the 80% UCL approach (points represent pilot trial sample sizes of 20, 24, 30 and 40) and Line 4, Main trial sample size calculation based on the 95% UCL approach (point represents pilot trial sample size of 55).

Table 6 (80% powered main trial and 90% powered main trial) represents the optimal results with a floor on the lower limit of the pilot study sample size at 10 per treatment group.

It should also be noted that although the exact calculation for the NCT approach (equation 4) has been used here to gain the most accurate results, in practice using the approximation in equation 5 will result in an overall study sample size of one subject less than the exact calculation at the most.

Table 6 shows again that the NCT method produces smaller overall trial sample sizes than both the 80% and 95% UCL methods. There is, on average, no loss of power when using the NCT approach, simulations and the results can be seen in Table 7. In order to calculate the results in Table 7, a pilot trial was simulated with two treatment arms. The results were drawn from a normal distribution, the control arm with a mean of 0 and a variance of 1 and the experimental arm with a mean of the required effect size and variance of 1. Depending on the adjustment method we were looking at, the pilot trial was set to the optimal sample size for that approach and effect size. The standard deviation was estimated from the pilot trial, this was then used to calculate the sample size

Table 7. Average power for two armed trials designed using different adjustment methods based on 10,000 simulations using 90% power, 5% Type I error rate and 'optimal' pilot trial sample sizes.

Standardised effect size		80% upper confidence limit	95% upper confidence limit	Non-central t-distribution
0.05	Pilot trial sample size	506	796	212
	Average power	91.25	92.51	90.52
	Percentage of trials with power above 90%	81.71	95.19	57.91
	Percentage of trials with power above 80%	100.00	100.00	99.87
0.1	Pilot trial sample size	212	334	108
	Average power	92.12	93.21	90.34
	Percentage of trials with power above 90%	82.15	95.93	60.34
	Percentage of trials with power above 80%	99.98	100.00	99.00
0.2	Pilot trial sample size	92	144	56
	Average power	93.17	94.75	90.36
	Percentage of trials with power above 90%	83.12	95.87	64.2
	Percentage of trials with power above 80%	99.74	100.00	96.15
0.5	Pilot trial sample size	32	52	24
	Average power	94.37	96.66	92.09
	Percentage of trials with power above 90%	84.19	95.60	68.90
	Percentage of trials with power above 80%	97.89	99.86	91.00
0.8	Pilot trial sample size	20	32	20
	Average power	95.37	97.60	92.10
	Percentage of trials with power above 90%	84.73	95.85	69.45
	Percentage of trials with power above 80%	95.33	99.53	89.23

for the main trial (the method depending on the approach under investigation). The main trial sample size calculations were based on a Type I error rate of 5%, a Type II error rate of 10% and an allocation ratio between the treatment groups of 1. Using the same method as with the pilot trial, the main trial was then simulated based on this sample size. The results of the main trial were then tested using a *t*-test. This simulation was repeated 10,000 times for each situation. The analysis was carried out in R 3.1.2.

From the simulations, the NCT approach gives the simulated average power closest to the nominal power level. When the standardised effect size is large, the 95% UCL approach has an average power approximately 7% above the nominal value.

4.2 Rules of thumb revisited

In many trials, the actual value of standardised effect size to be used in the main trial may not be known before the pilot trial planning stage. This is one of the reasons that the existing rules of thumb for the pilot trial sample size, as introduced earlier in the paper, are so attractive. However, an investigator is likely to know whether the standardised difference for use in the main trial is likely to be small, medium or large within a range.

From the results presented, it would seem that any rules of thumb should be stepped – and not flat – so that the pilot is bigger for smaller standardised effect sizes and smaller for large standardised effect sizes.

Table 8. Estimated stepped rules of thumb for required pilot trial sample size per treatment arm when the NCT approach will be used to calculate the main trial sample size.

Standardised difference	80% powered main trial	90% powered main trial
Extra small ($\delta < 0.1$)	50	75
Small ($0.1 \leq \delta < 0.3$)	20	25
Medium ($0.3 \leq \delta < 0.7$)	10	15
Large ($\delta \geq 0.7$)	10	10

Table 6 (80% powered main trial and 90% powered main trial) has been used to derive new stepped rules of thumb for the pilot trial sample size; these are presented in Table 8. These offer (per arm) sample sizes for pilot trials, which vary depending on whether the standardised effect size for the main trial is small, medium or large. An additional category of extra small has been inserted into Cohen's classifications, which represents standardised effect sizes of 0.1 or less; this is because the results for these trials were many times larger than for standardised effect sizes of 0.2.

5 A worked example

A two armed parallel group randomised controlled clinical trial is being planned with a two-sided Type I error rate of 5% and a power of 90%. The primary outcome is anticipated to take a normal form. As the investigator initially was unsure about design aspects of the main trial such as the anticipated standard deviation of the outcome measure and the likely recruitment and dropout rates, a pilot trial was undertaken.

Initially a flat rule of thumb was used, and the pilot sample size was chosen to be 24 evaluable patients in total as recommended by Julious.[16]

However, suppose that *a priori* the standardised effect size for the main trial is 0.25. Using the NCT approach, the main trial sample size is estimated to be 760 participants, assuming that the pilot trial of 24 was used to design the trial. This would result in a total sample size for the pilot and main trial together of 784 participants.

As highlighted previously, if the standardised effect size to be used in the main trial is known to be 0.25 prior to the pilot trial, then based on the method presented in this paper, the optimal pilot trial sample size for a standardised difference of 0.25 is 46. If a pilot trial of 46 participants was carried out and the main trial planned based on the estimate of the standard deviation from that pilot study; then the main trial sample size based on the NCT approach would be 716. This method would result in a total overall sample size of 762 participants.

Thus, by increasing the sample size for the pilot trial, in this example nearly doubling the sample size, we have increased the precision around the standard deviation estimate. This has had the effect of reducing the total trial sample size by 22.

There are many instances where the effect size for the main trial is unlikely to be known prior to the pilot trial. However, it could be considered reasonable to have an approximate idea of the sample size of the main trial based on experience of the same population, i.e. it is anticipated that the effect size will be quite small and the sample size large. Using the stepped rules of thumb (from Table 8), the sample size would be set at 50 for the pilot trial. Consequently, the main trial sample size calculation based on a standardised effect size of 0.25 would be for 712 subjects; giving a total overall trial sample size of 762. In this example due to rounding, the total sample size is the same for the stepped rules of thumb approach and the optimal solution.

6 Discussion

The National Institute for Health Research Evaluation, Trials and Studies Coordinating Centre defines pilot trials in context of the planning of a future trial.[20] Therefore, the method of minimising the sample size across trials could be thought to be the most appropriate as it treats the pilot trial as part of the whole study programme rather than a stand-alone trial. In this paper, we propose a method for estimating the sample size for a pilot trial, which uses this idea. The method introduced describes how to set the sample size of a pilot trial in order to minimise the overall trial sample size, i.e. the sample size of the pilot and main trial together, for different correction methods.

We demonstrate how the size of the pilot trial impacts on the size of the overall trial when either the UCL approach or the NCT method is used to calculate the sample size for the main trial. If the pilot trial is large, the main trial will be relatively small and if the pilot trial is small, the main trial will be relatively large. It can be seen from the results in this paper that the NCT approach provides lower overall trial sample sizes than any other method while maintaining the average power at the nominal level.

Our results show that as the sample size of a main trial increases, the size of the pilot trial should also increase. For medium effect sizes, the existing rules seem sufficient; however, as we move away from a standardised effect size of 0.5, the flat rules of thumb can over or under estimate the pilot trial sample size that would minimise the overall trial sample size. Therefore, using these flat rules of thumb would lead to more patients than theoretically required being recruited to the overall trial. This is especially seen at small standardised effect sizes.

From the results presented in this paper, we recommend using the NCT approach to set the main trial sample size in conjunction with the method presented of calculating a pilot trial sample size. Doing so will on average maintain the nominal power requirement and minimise the overall trial sample size for the pilot and the main trial together.

If simpler calculations are to be undertaken for a pilot trial sample size, we recommend using the stepped rules of thumb presented in the paper to set the pilot study sample size. However, if the standardised effect size to be used in the main trial is known, we recommend that the exact calculation be used.

In the paper, the emphasis is on estimating the sample size for pilot trials to minimize the overall sample size across both the main and pilot trial combined. However, there could be other sample size considerations such as obtaining plausible estimates of the clinical effect through precision of the confidence intervals.[21–25] Alternatively, decision science criteria could be used to optimize the risk discharged in a clinical development prior to the start of a late phase study.[26] In both these instances, a pilot trial is still considered in context with later definitive trials but there may already – from previous work – be good estimates of the population variance.

Finally, the methods described in the paper do have limitations. The main assumption is that the design of the main trial and the pilot trial is ostensibly the same. This may not be the case, however, which could impact on the applicability of the estimate of the standard deviation from the pilot trial.

The methods described in the paper provide a way to estimate the optimal pilot trial sample size that minimises the overall sample size for a given main trial standardised effect size. We recognise that the situation of knowing the effect size prior to the pilot study is an ideal situation and so we recommend that the stepped rules of thumb, proposed in this paper, be used and the flat rules of thumb only used as a last resort.

Acknowledgements

We would like to thank the reviewer for their considered and valuable comments, which have vastly improved this manuscript.

Funding

ALW, SAJ, CLC and MJC are funded by The University of Sheffield.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Campbell MJ, Machin D and Walters SJ. *Medical statistics: a textbook for the health sciences*. Chicester, UK: John Wiley & Sons Inc, 2010.
2. Halpern SD, Karlawish JHT and Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002; **288**(3): 358–362.
3. Machin D, Campbell MJ, Tan SB, et al. *Sample size tables for clinical studies*, 3rd ed. Chichester, UK: Wiley-Blackwell, 2008.
4. Julious SA. *Sample sizes for clinical trials*. Florida, USA: Chapman and Hall, 2009.
5. Whitehead AL, Sully BGO and Campbell MJ. Pilot and feasibility studies: is there a difference from each other and from a randomised controlled trial? *Contemp Clin Trials* 2014; **38**: 130–133.
6. Pocock SJ. *Clinical trials: a practical approach*. Chicester, UK: John Wiley & Sons Ltd, 1983.
7. Friede T and Kieser M. A comparison of methods for adaptive sample size adjustment. *Stat Med* 2001; **20**: 3861–3873.
8. Denne JS and Jennison C. Estimating the sample size for a T-test using an internal pilot. *Stat Med* 1999; **18**: 1575–1585.
9. Julious SA and Owen RJ. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceut Stat* 2006; **5**: 29–37.
10. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995; **14**: 1933–1940.
11. Sim J and Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *J Clin Epidemiol* 2012; **65**: 301–308.
12. Kieser M and Wassmer G. On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biom J* 1996; **8**: 941–949.
13. Lancaster GA, Dodd S and Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract* 2004; **10**: 307–312.
14. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol* 2010; **10**: 1.
15. Arain M, Campbell MJ, Cooper CL, et al. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol* 2010; **10**: 67.
16. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceut Stat* 2005; **4**: 287–291.
17. Teare MD, Dimairo M, Shephard N, et al. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014; **15**: 264.
18. Cohen J. A power primer. *Psychol Bull* 1992; **112**: 155–159.
19. Birkett MA and Day SJ. Internal pilot studies for estimating sample size. *Stat Med* 1994; **13**: 2455–2463.
20. NETSCC. Glossary: feasibility and pilot studies, http://www.netscc.ac.uk/glossary/ (accessed 8 October 2012).
21. Day S. Clinical trial numbers and confidence intervals of pre-specified size. *Lancet* 1988; **332**: 1427.
22. Julious SA and Patterson SD. Sample sizes for estimation in clinical research. *Pharmaceut Stat* 2004; **3**: 213–215.
23. Grieve AP. Sample sizes and confidence intervals. *Am Stat* 1990; **44**: 190.
24. Wood J and Lambert M. Sample size calculations for trials in health services research. *J Health Serv Res Policy* 1999; **4**: 226–229.
25. Lee EC, Whitehead AL, Jacques RM, et al. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Med Res Methodol* 2014; **14**: 41.
26. Julious SA and Swank DJ. Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan. *Pharmaceut Stat* 2005; **4**: 37–46.

Pilot and feasibility studies: Is there a difference from each other and from a randomised controlled trial?

Amy L. Whitehead MSc, Benjamin G.O. Sully MSc, Michael J. Campbell PhD[§]

Address:

Design, Trials and Statistics Group,

School of Health and Related Research,

University of Sheffield,

Regent Court,

30 Regent Street,

Sheffield, UK

S1 4DA.

[§]Corresponding Author: Michael J Campbell

Email: m.j.campbell@sheffield.ac.uk +44 114 222 0839

Abstract

Background: A crucial part in the development of any intervention is the preliminary work carried out prior to a large scale definitive trial. However, the definitions of these terms are not clear cut and many authors redefine them. Because of this, the terms *feasibility* and *pilot* are often misused.

Aim: To provide an introduction to the topic area of pilot and feasibility trials and draw together the work of others in the area on defining what is a pilot or feasibility study.

Methods: A review of definitions and advice from the published literature and from funders' websites. Examples are used to show evidence of good practice and poor practice.

Results: We found that researchers use different terms to describe the various stages of the research process. Some define the terms feasibility and pilot as being different whereas others argue that these terms are synonymous. All reflective papers agree that feasibility/pilot studies should not test treatment comparisons nor estimate feasible effect sizes. However, this is not universally observed in practice.

Summary: We believe that the term 'feasibility' should be used as an overarching term for preliminary studies and the term 'pilot' refers to a specific type of study which resembles the intended trial in aspects such as, having a control group and randomisation. However, studies labelled 'pilot' should have different aims and objectives to main trials, and an intention for further work in the future. Researchers should not use the title 'pilot' for a trial which evaluates a treatment effect.

Keywords: Pilot; feasibility; terminology; reporting;

Introduction

During recent years there has also been an increasing emphasis on the importance of preliminary work prior to the organisation of large-scale, publicly-funded randomised controlled trials. Many large public funding bodies now expect substantial work to have been done prior to the main bid. Some funding streams, such as UK National Institute for Health Research (NIHR) Research for Patient Benefit (RfPB) [1], and the US NIH R34 funding mechanism [2] recognise this through the provision of substantial sums of money to support such work. The value of preliminary work is now recognised and researchers are encouraged to publish their pilot work in advance of their main trial, and some publishers are willing to publish such results. However there remains much confusion about the purpose of preliminary work and also of terminology used. The NIHR use the terms 'feasibility' and 'pilot' to distinguish between different stages in the research process [3]. Although these terms are frequently used in the literature they are used inconsistently and interchangeably[4]; while other authors choose to use different terms completely to define the stages of development [5].

There is also the temptation to label a trial 'pilot' to excuse a small sample size, or one conducted in one locality, but still with the intention of running a study with treatment comparison as the main objective.

The aim of this paper is to provide an introduction to the topic area of pilot and feasibility trials. We will draw together the work of others that has been done in this area, describing current definitions, their overlaps and points of divergence. We use examples to illustrate good and poor practice and conclude with some recommendations on the use of the terms. This paper adds to our earlier work[4] by critiquing earlier definitions, and providing examples to support our criticism.

Current Definitions

Within the pharmaceutical sector testing drug efficacy has long had a tradition of clearly defined stages, from the initial phase 1, first-into-man studies through to the phase 4 post-marketing studies. However, for large publicly funded trials, particularly of complex interventions and modes of care, the definitions and stages of trials have been less well defined/clear-cut. There have been several attempts to provide guidance on the definitions of a pilot and feasibility study. A review of papers published in 2001 in seven major journals looked at the objectives of pilot studies in the literature [6] to clarify the definition of pilot study. This was repeated in 2010 and the work extended to distinguish between pilot and feasibility studies in the article search and looking at the components of the studies [4]. The authors of these studies found that studies labelled 'pilot' generally used stricter methodology than studies labelled 'feasibility' and that pilot studies mostly reported their results as inconclusive and suggested further work, whereas feasibility studies did not state the same intention. They argue that the distinction between the two terms is not clear cut. However, they suggest the adoption of the NETSCC (NIHR Evaluation, Trials and Studies Coordinating Centre) definition which does distinguish between the two types of study [3].

The NETSCC [3] define feasibility studies as studies used to estimate important parameters that are needed to design the main study, e.g. standard deviation of the outcome measure, willingness of patients to be randomised, willingness of clinicians to recruit participants, number of people eligible, follow-up rates, response rates and adherence/ compliance rates. Feasibility studies may have no plan for further work and their aim is to assess whether it is possible to perform a full-scale study.

The NETSCC [3] define a pilot study as a version of the main study run in miniature to determine whether the components of the main study can all work together. They suggest that a pilot should focus on the processes of running the main study i.e. to ensure the

mechanisms of recruitment, randomisation, treatment and follow-up assessments. The aim of the pilot is to provide training and experience in the running of the trial and to highlight any problems, so they may be corrected before the main study begins. There must also be a plan for further work. A pilot study can be either external or internal to the main study.

This latter definition is comparable to the UK NICE definition of a pilot study as “a small-scale ‘test’ of a particular approach … The aim would be to highlight any problems or areas of concern and amend it before the full-scale study begins.” [7]

However, in contrast Arnold et al [5] provided three separate definitions for different types of pre-clinical work: pilot work, pilot studies and pilot trials. They defined pilot *work* as “any background research that informs a future study”; pilot *studies* as “studies with a specific hypothesis, objective and methodology”; and a pilot *trial* as “a stand-alone pilot study with a randomisation procedure”. Indeed the authors advocated against using the term *feasibility study*, arguing that it “does not reflect the scope of many pilot studies”. These definitions differ from most others in that they distinguish between the different possible objectives of pilot studies, but do not include the term feasibility whatsoever. The movement through development stages is defined by using the words; work, study and trial instead of the terms feasibility and pilot.

Thabane et al. [8], in their tutorial on pilot studies, do not distinguish between feasibility and pilot studies, and simply note that the terms are used synonymously. They do however note that the main focus of a pilot study should be to test the feasibility of conducting a full study, rather than statistical significance, and that many pilot studies fail to recognise this.
Leon et al [9] state that a pilot study can be used to evaluate the feasibility of recruitment, randomization, retention, assessment procedures, new methods, and implementation of the novel intervention. A pilot study is not a hypothesis testing study. Safety, efficacy and effectiveness are not evaluated in a pilot. Contrary to tradition, a pilot study does not provide

a meaningful effect size estimate for planning subsequent studies due to the imprecision inherent in data from small samples. Thus effect sizes provided by pilot studies should not be used to power a subsequent full trial. Instead clinical experience should be used to define a *clinically meaningful* effect. A pilot study is a requisite initial step in exploring a novel intervention or an innovative application of an intervention. Pilot results can inform feasibility and identify modifications needed in the design of a larger, ensuing hypothesis testing study.

This is similar to the British Medical Research Council's (MRC's) complex interventions guidelines which urge the reader to exercise caution when using the results of a pilot study to make assumptions about the required sample size, likely response rates etc., when the evaluation is scaled up [10]. These guidelines do not give an exact definition of a pilot or feasibility study, instead focusing on the outcomes of the feasibility and piloting stage. Investigators should be confident that the intervention can be delivered as intended and be able to make safe assumptions about the effect sizes, variability, recruitment rates and retention to aid in the designing of the main study. They do note that "a pilot study need not be a 'scale model' of the planned main stage evaluation, but should address the main uncertainties that have been identified in the development work".

Examples

Krarup et al [11] describe a trial, the ExSTroke Pilot trial, to examine the benefits of exercise in patients who have had a stroke. They intended to recruit 300 subjects, but this was powered on a postulated difference in treatment groups from a surrogate outcome, the Physical Activity Scale for the Elderly (PACE). The reason for the term 'pilot' in the title could be inferred because the study was not powered for recurrent stroke, MI, or mortality. The results were published [12] as a randomised controlled trial. The trial was criticised because it did not follow guidelines for the developing of complex interventions such as those of the MRC [10] and "we might have expected modelling of active ingredients of the intervention (given that it was a pilot study) and testing the feasibility of the approach" [13].

In contrast, the LIFE study [14] is also described as a pilot study. The study intended to recruit 400 adults and the aims were: (a) estimate the sample size needed for a full scale trial, (b) examine the consistency of the effects of the physical activity intervention on several continuous measures of physical function, (c) assess the feasibility of recruitment, (d) evaluate study adherence and retention, (d) evaluate the efficacy of a stepped care approach for managing intercurrent illness in this at-risk population, and (e) develop a comprehensive system for monitoring and ensuring participant safety. Two points can be made, firstly the objectives of the study are consistent with the objectives of a pilot study, except (d) since it was not powered to evaluate efficacy. Secondly the size of the projected pilot, at 400, exceeds the size of many full studies and is not justified in relation to the objectives. The outcomes of some of these objectives were subsequently published. For example the investigators evaluated the longitudinal distributions of four standardized outcomes to contrast how they may serve as primary outcomes of future clinical trials: ability to walk 400 meters, ability to walk 4 meters in ≤10 seconds, a physical performance battery, and a questionnaire focused on physical function. They concluded that the ability to walk 400 meters as a dichotomous outcome provided the smallest sample size projections and that a 4-year trial based on the outcome of the 400-meter walk is projected to require *n*= 962–2234 to detect an intervention effect of 30%–20% with 90% power[15]. In fact they are now running the main study, a trial of 1600 people followed up for 2.7 years [16]. This outcome is entirely coherent with that of the pilot study. However, in view of the size of the pilot, they could not resist also doing some treatment comparisons [17-18]! It is also of note that the size of the pilot was 25% of the main study, which leads one to query the correct ratio in size of the pilot and main study.

Discussion

It can be seen that there is still confusion around the terms. Some use the terms feasibility and pilot interchangeably [8] whereas others define the terms separately [3,4,6]. It is problematic to look to the literature to find a difference between pilot and feasibility study as

a trial may be labelled as a pilot or feasibility study but this doesn't mean that it is a pilot or feasibility study under someone else's definition.

From the review of the literature we found that the distinguishing features of a pilot study from a feasibility study are:

- Stricter study methodology (e.g. a justification of the sample size)
- An intention for further work
- Smaller version of the main study
- A focus on trial processes

The stricter methodology may stem from the fact that pilot studies are more likely to mimic the design of the main study, in order to test the processes and provide training to trial staff and alleviate problems before the larger trial. This restriction does not hold for a feasibility study, where a systematic review or meta-analysis may be a feasibility study. A pilot study, apart from investigating how the trial procedures will work in the future trial, may also test the feasibility of a larger study so it could be said that pilot studies are also feasibility studies. However, the inverse cannot be said; that all feasibility studies are pilot studies. From this one could conclude that a pilot study is a special type of feasibility study which has a plan for further work and mimics the envisioned definitive trial. In addition, we could also define a pilot *trial* as a pilot study which also involves randomisation between treatment groups.

The plan for further work is crucial for pilot studies otherwise the study may be seen as an underpowered trial which are deemed unethical and have limited scientific use. As we have shown pilot studies and randomised controlled trials (RCTs) have different aims and objectives [4]. An RCT will test the efficacy of a new intervention, a pilot study should only test other aspects of the trial design in preparation for this definitive assessment of the treatment. The term 'pilot' implies an intention for further definitive work in the future.

It is impossible to legislate on the use of terminology, but we suggest that if journals and reviewers adopt a more consistent usage, then it would make the reporting and reviewing of such studies much simpler.

It could be argued that trials which use a surrogate endpoint, such as the ExStroke trial [11] are in fact 'pilot' studies even if they test for treatment comparisons. However, to be consistent with the previous paragraph, they only deserve this label if there are clear criteria to decide on a trial using clinically meaningful outcomes, and a clear intention of conducting such a trial if the criteria are met. Otherwise the title should clearly define the trial as one that uses surrogate endpoints. Thus the ExStroke trial could have specified what size difference in the PACE outcome would have justified further follow up for stroke and death, or an extension of the trial to include these outcomes.

Conclusion

The distinction between pilot and feasibility studies is still a very grey area, with various definitions having been suggested by clinical trial methodology researchers. We suggest it is futile to ascribe a particular meaning to the term 'feasibility' and that all preliminary trial work could be described as 'feasibility' therefore it could be thought of as an overarching term for preliminary work. However the term 'pilot' could be reserved for a study that mimics the definitive trial design in that it may include control groups and randomisation, but whose explicit objective is *not* to compare treatment groups, but rather to ensure the main trial delivers maximum benefit. Trials that use surrogate endpoints could be described as pilot trials only if they include clear criteria for proceeding to a main trial.

Acknowledgements

We thank members of the CONSORT team on Pilot and Feasibility trials for helpful discussion (Sandra Eldridge, Gill Lancaster, Christine Bond, Sally Hopewell and Lehana Thabane)

References

1. NIHR Research for Patient Benefit Programme. Director's Message 6 - Thinking about applying for funding for a pilot study? (accessed December 18 2013) http://www.ccf.nihr.ac.uk/RfPB/Documents/RfPB_Directors_message_6.pdf.

2. National Institute of Mental Health. Pilot Intervention and Services Research Grants (R34). (accessed December 18 2013). http://grants.nih.gov/grants/guide/pa-files/PAR-09-173.html

3. NETSCC. Glossary: Feasibility and Pilot Studies. 2013 (accessed December 18 2013) http://www.netscc.ahttp://www.nets.nihr.ac.uk/glossary?result_1655_result_page=Pc.uk/glossary/

4. Arain, M., Campbell MJ, Cooper CL and Lancaster GA What is a Pilot or Feasibility Study? A Review of Current Practice and Editorial Policy. *BMC Medical Research*

Methodology, 2010. 10: 67.

5. Arnold DM., Burns KE, Adhikari NK, Kho ME, Meade MO, Cook DJ, The Design and Interpretation of Pilot Trials in Clinical Research in Critical Care. *Critical Care Medicine*, 2009. 37(1): p. S69-S74.

6. Lancaster GA., Dodd S, and Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 2004. 10(2): 307-312.

7. NICE. Glossary. (accessed December 19 2013) http://www.nice.org.uk/website/glossary/glossary.jsp?alpha=P.

8. Thabane, L., Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Goldsmith CH l., A Tutorial on Pilot Studies: The What, Why and How. *BMC Medical Research Methodology,* 2010. 10.

9. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatry Res*, 2011, 45(5), 626-629.

10. Medical Research Council. Developing and Evaluating Complex Interventions: New guidance. (accessed December 18 2013) http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC004871

11. Krarup L-H, Gluud C, Truelsen T, Pedersen A, Lindahl M, et al. The ExSTroke Pilot Trial: rationale, design, and baseline data of a randomized multicenter trial comparing physical training versus usual care after an ischemic stroke. *Contemporary Clinical Trials* 29, 3 (2008): 410-417.

12. Boysen G, Krarup L-H, Zeng X, Oskedra A, Kõrv J, et al. ExStroke Pilot Trial of the effect of repeated instructions to improve physical activity after ischaemic stroke: a multinational randomised controlled clinical trial. *BMJ* 2009; 339:b2810

13**.** Mutrie N. ExStroke Pilot Trial of the effect of repeated instructions to improve physical activity after ischaemic stroke: a multinational randomised controlled clinical trial: *Responses Tab* .http://www.bmj.com/content/339/bmj.b2810?tab=responses (accessed 19th Dec 2013)

14. Rejeski, WJ, Fielding RA, Blair SB, Guralnik JM, Gill TM, et al. The lifestyle

interventions and independence for elders (LIFE) pilot study: design and methods. *Contemporary Clinical Trials* 2005; 26: 141-154.

15. Espeland, M. A., Gill, T. M., Guralnik, J., Miller, M. E., Fielding, R., Newman, A. B., & Pahor, M. (2007). Designing clinical trials of interventions for mobility disability: results from the lifestyle interventions and independence for elders pilot (LIFE-P) trial. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*,*62*(11), 1237-1243

16. The Life Study. General Public Information. 2013 (accessed 19th Dec 2013) https://www.thelifestudy.org/public/index.cfm

17. Pahor M, Blair SN, Espeland M, Fielding R, Gill TM, Guralnik JM et al. Effects of a physical activity intervention on measures of physical performance: Results of the lifestyle interventions and independence for Elders Pilot (LIFE-P) study. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 2006;61(11):1157-65..

18. Rejeski WJ, Marsh AP, Chmelo E, Prescott AJ, Dobrosielski M, et al. (2009). The lifestyle interventions and independence for elders pilot (LIFE-P): 2-year follow-up. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *64*(4), 462-467.

Teare et al. Trials 2014, 15:264
http://www.trialsjournal.com/content/15/1/264

TRIALS

RESEARCH Open Access

Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study

M Dawn Teare*, Munyaradzi Dimairo, Neil Shephard, Alex Hayman, Amy Whitehead and Stephen J Walters

Abstract

Background: External pilot or feasibility studies can be used to estimate key unknown parameters to inform the design of the definitive randomised controlled trial (RCT). However, there is little consensus on how large pilot studies need to be, and some suggest inflating estimates to adjust for the lack of precision when planning the definitive RCT.

Methods: We use a simulation approach to illustrate the sampling distribution of the standard deviation for continuous outcomes and the event rate for binary outcomes. We present the impact of increasing the pilot sample size on the precision and bias of these estimates, and predicted power under three realistic scenarios. We also illustrate the consequences of using a confidence interval argument to inflate estimates so the required power is achieved with a pre-specified level of confidence. We limit our attention to external pilot and feasibility studies prior to a two-parallel-balanced-group superiority RCT.

Results: For normally distributed outcomes, the relative gain in precision of the pooled standard deviation (SD_p) is less than 10% (for each five subjects added per group) once the total sample size is 70. For true proportions between 0.1 and 0.5, we find the gain in precision for each five subjects added to the pilot sample is less than 5% once the sample size is 60. Adjusting the required sample sizes for the imprecision in the pilot study estimates can result in excessively large definitive RCTs and also requires a pilot sample size of 60 to 90 for the true effect sizes considered here.

Conclusions: We recommend that an external pilot study has at least 70 measured subjects (35 per group) when estimating the SD_p for a continuous outcome. If the event rate in an intervention group needs to be estimated by the pilot then a total of 60 to 100 subjects is required. Hence if the primary outcome is binary a total of at least 120 subjects (60 in each group) may be required in the pilot trial. It is very much more efficient to use a larger pilot study, than to guard against the lack of precision by using inflated estimates.

Keywords: sample size, feasibility studies, pilot studies, binary outcomes, continuous outcomes, RCTs

Background

In 2012/13, the National Institute for Health Research (NIHR) funded £208.9 million of research grants across a broad range of programmes and initiatives to ensure that patients and the public benefit from the most cost-effective up-to-date health interventions and treatments as quickly as possible [1]. A substantial proportion of these research grants were randomised controlled trials (RCTs) to assess the clinical effectiveness and cost-effectiveness of new health technologies. Well-designed RCTs are widely regarded as the least biased research design for evaluating new health technologies and decision-makers, such as the National Institute for Health and Care Excellence (NICE), are increasingly looking to the results of RCTs to guide practice and policy.

RCTs aim to provide precise estimates of treatment effects and therefore need to be well designed to have good power to answer specific clinically important questions. Both overpowered and underpowered trials are

* Correspondence: m.d.teare@sheffield.ac.uk
Design, Trials and Statistics Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, S1 4DA Sheffield, UK

undesirable and each poses different ethical, statistical and practical problems. Good trial design requires the magnitude of the clinically important effect size to be stated in advance. However, some knowledge of the population variation of the outcome or the event rate in the control group is necessary before a robust sample size calculation can be done. If the outcome is well established, these key population or control parameters can be estimated from previous studies (RCTs or cohort studies) or through meta-analyses. However, in some cases finding robust estimates can pose quite a challenge if reliable data, for the proposed trial population under investigation, do not already exist.

A systematic review of published RCTs with continuous outcomes found evidence that the population variation was underestimated (in 80% of reported endpoints) in the sample size calculations compared to the variation observed when the trial was completed [2]. This study also found that 25% of studies were vastly underpowered and would have needed five times the sample size if the variation observed in the trial had been used in the sample size calculation. A more recent review of trials with both binary and continuous outcomes [3] found that there was a 50% chance of underestimating key parameters. However, they too found large differences between the estimates used in the sample size calculation compared to the estimates derived from the definitive trial. This suggests that many RCTs are indeed substantially underpowered or overpowered. A systematic review of RCT proposals reaching research ethics committees [4] found more than half of the studies included did not report the basis for the assumed values of the population parameters. So the values assumed for the key population parameters may be the weakest part of the RCT design.

A frequently reported problem with publicly funded RCTs is that the recruitment of participants is often slower or more difficult than expected, with many trials failing to reach their planned sample size within the originally envisaged trial timescale and trial-funding envelope. A review of a cohort of 122 trials funded by the United Kingdom (UK) Medical Research Council and the NIHR Health Technology Assessment programme found that less than a third (31%) of the trials achieved their original patient recruitment target, 55/122 (45.1%) achieved less than 80% of their original target and half (53%) were awarded an extension [5]. Similar findings were reported in a recently updated review [6]. Thus, many trials appear to have unrealistic recruitment rates. Trials that do not recruit to the target sample size within the time frame allowed will have reduced power to detect the pre-specified target effect size.

Thus the success of definitive RCTs is mainly dependent on the availability of robust information to inform the design. A well-designed, conducted and analysed pilot or feasibility trial can help inform the design of the definitive trial and increase the likelihood of the definitive trial achieving its aims and objectives. There is some confusion about terminology and what is a feasibility study and what is a pilot study. UK public funding bodies within the NIHR portfolio have agreed definitions for pilot and feasibility studies [7]. Other authors have argued against the use of the term 'feasibility' and distinguish three types of preclinical trial work [8].

Distinguishing features of pilot and feasibility studies

NIHR guidance states:

Feasibility studies are pieces of research done before a main study in order to answer the question 'Can this study be done?'. In this context they can be used to estimate important parameters that are needed to design the main study [9]. *For instance:*

i) *standard deviation of the outcome measure, which is needed in some cases to estimate sample size;*
ii) *willingness of participants to be randomised;*
iii) *willingness of clinicians to recruit participants;*
iv) *number of eligible patients over a specific time frame;*
v) *characteristics of the proposed outcome measure and in some cases feasibility studies might involve designing a suitable outcome measure;*
vi) *follow-up rates, response rates to questionnaires, adherence/compliance rates, intracluster correlation coefficients in cluster trials, etc.*

Feasibility studies for randomised controlled trials may themselves not be randomised. Crucially, feasibility studies do not evaluate the outcome of interest; that is left to the main study.

If a feasibility study is a small RCT, it need not have a primary outcome and the usual sort of power calculation is not normally undertaken. Instead the sample size should be adequate to estimate the critical parameters (e.g. recruitment rate) to the necessary degree of precision.

Pilot trials are a version of the main study that is run in miniature to test whether the components of the main study can all work together [9]. *It will therefore resemble the main study in many respects, including an assessment of the primary outcome. In some cases this will be the first phase of the substantive study and data from the pilot phase may contribute to the final analysis; referred to as an internal pilot. Or at the end of the pilot study the data may be analysed and set aside, a so-called external pilot* [10].

For the purposes of this paper we will use the term pilot study to refer to the pilot work conducted to estimate key parameters for the design of the definitive trial.

There is extensive but separate literature on two-stage RCT designs using an internal pilot study [11-14].

There is disagreement over what sample size should be used for pilot trials to inform the design of definitive RCTs [15-18]. Some recommendations have been developed although there is no consensus on the matter. Furthermore, the majority of the recommendations focus on estimating the variability of a continuous outcome and relatively little attention is paid to binary outcomes. The disagreement stems from two competing pressures. Small studies can be imprecise and biased (as defined here by comparing the median of the sampling distribution to the true population value), so larger sample sizes are required to reduce both the magnitude of the bias and the imprecision. However, in general participants measured in an external pilot or feasibility trial do not contribute to the estimation of the treatment effect in the final trial, so our aim should be to maintain adequate power while keeping the total number of subjects studied to a minimum. Recently some authors have promoted the practice of taking account of the imprecision in the estimate of the variance for a continuous outcome. Several suggest the use of a one-sided confidence interval approach to guarantee that power is at least what is required more than 50% of the time [15,18,19].

This paper aims to provide recommendations and guidelines with respect to two considerations. Firstly, what is the number of subjects required in an external pilot RCT to estimate the uncertain critical parameters (SD for continuous outcomes; and consent rates, event rates and attrition rates for binary outcomes) needed to inform the design of the definitive RCT with a reasonable degree of precision? Secondly, how should these estimates from the pilot study be used to inform the sample size (and design) for the definitive RCT? We shall assume that the pilot study (and the definitive RCT) is a two-parallel-balanced-group superiority trial of a new treatment versus control.

For the purposes of this work we assume that the sample size of the definitive RCT is calculated using a level of significance and power argument. This is the approach that is currently commonly employed in RCTs; however, alternative methods to calculate sample size have been proposed, such as using the width of confidence intervals [20] and Bayesian approaches to allow for uncertainty [21-23].

Methods

Our aim is to demonstrate the variation in estimates of population parameters taken from small studies. Though the sampling distributions of these parameters are well understood from statistical theory, we have chosen to present the behaviours of the distributions through simulation rather than through the theoretical arguments as the visual representation of the resulting distributions makes the results accessible to a wider audience.

Randomisation is not a necessary condition for estimating all parameters of interest. However, it should be noted that some parameters of interest during the feasibility phase are related to the randomisation procedure itself, such as the rate of willingness to be randomised, and the rate of retention or dropout in each randomised arm. In addition, randomisation ensures the equal distribution of known and unknown covariates on average across the randomised groups. This ensures that we can estimate parameters within arms without the need to worry about confounding factors. In this work we therefore decided to allow for the randomisation of participants to mimic the general setting for estimating all parameters, although it is acknowledged that some parameters are independent of randomisation.

We first consider a normally distributed outcome measured in two groups of equal size. We considered study groups of from 10 to 80 subjects using increments of five per group. For each pilot study size, 10,000 simulations were performed. Without loss of generality, we assumed the true population mean of the outcome is 0 and the true population variance is 1 (and that these are the same in the intervention and control groups). We then use the estimate of the SD, along with other information, such as the minimum clinically important difference in outcomes between groups, and Type I and Type II errors levels, to calculate the required sample size (using the significance thresholds approach) for the definitive RCT.

The target difference or effect size that is regarded as the minimum clinically important difference is usually the difference in the means when comparing continuous outcomes for the intervention with those of the control group. This difference is then converted to a standardised effect size by dividing by the population SD. More details of the statistical hypothesis testing framework in RCTs can be found in the literature [24,25].

For a two-group pilot RCT we can use the SD estimate from the new treatment group or the control/usual care group or combine the two SD estimates from the two groups and use a pooled standard deviation (SD_p) estimated from the two-group specific sample SDs. For sample size calculations, we generally assume the variability of the outcome is the same or equal in both groups, although this assumption can be relaxed and methods are available for calculating sample sizes assuming unequal SDs in each group [26,27]. This is analogous to using the standard *t*-test with two independent samples (or multiple linear regression), which assumes equal variances, to analyse the outcome data compared with using versions of the *t*-test that do not assume equal variances (e.g. Satterthwaite's or Welch's correction).

We assume binary outcomes are binomially distributed and consider a number of different true population proportions as the variation of proportion estimator is a function of the true proportion. When estimating an event rate, it may not always be appropriate to pool the two arms of the study so we study the impact of estimating a proportion from a single arm where the study size increases in steps of five subjects. We considered true proportions in the range 0.1 to 0.5 in increments of 0.05. For each scenario and sample size, we simulated the feasibility study at least 10,000 times depending on the assumed true proportion. For the binary outcomes, the number of simulations was determined by requiring the proportion to be estimated within a standard error of 0.001. Hence, the largest number of simulations required was 250,000 when the true proportion was equal to 0.5. Simulations were performed in Stata version 12.1 [28] and R version 13.2 [29].

Normally distributed outcomes

For each simulation, sample variances were calculated for each group (s_1^2 and s_2^2) and the pooled SD was calculated as follows:

$$\mathrm{SD}_p = \sqrt{\left(\frac{(s_1^2 + s_2^2)}{2}\right)}. \quad (1)$$

We also computed the standard error of the sample pooled SD which is

$$\mathrm{se}\left(\mathrm{SD}_p\right) = \frac{\mathrm{SD}_p}{\sqrt{2(n-1)}}. \quad (2)$$

To quantify the relative change in precision, we compared the average width of the 95% confidence intervals (WCI_{2n}) for the SD_p for study sizes of $2n$ with the average width when the study size was increased to $2(n + 5)$. We use the width of the confidence interval as this provides a measure of the precision of the estimate.

Given the sampling distribution of the SD, its lower and upper 95% confidence limits are given by:

$$\left(\sqrt{\frac{2(n-1)}{\chi_{0.025,2(n-1)}}}\mathrm{SD}_p \text{ and } \sqrt{\frac{2(n-1)}{\chi_{0.975,2(n-1)}}}\mathrm{SD}_p\right), \quad (3)$$

and the relative percentage gain in precision is quantified as the reduction in 95% confidence interval width if the sample size is increased by five per group:

$$\left(\frac{\mathrm{WCI}_{2n} - \mathrm{WCI}_{2(n+5)}}{\mathrm{WCI}_{2n}}\right) \times 100. \quad (4)$$

Bias is assessed by subtracting the true value from each estimate and taking the mean of these differences.

We also consider the impact of adjusting the SD estimate from the pilot as suggested originally by Browne in 1995 [15]. Here a one-sided confidence limit is proposed to give a corrected value. If we used the 50% one-sided confidence limit, this would adjust for the bias in the estimate, and this correction has also been proposed when using small pilots [17]. If we specify 50% confidence then our power will be as required 50% of the time. Sim and Lewis [18] suggest that it is reasonable to require that the sample size calculation guarantees the desired power with a specified level of confidence greater than 50%. For the sake of illustration, we will consider an 80% confidence level for the inflation factor. So we require the confidence interval limit associated with 80% confidence above that value. Hence the inflation factor to apply to the SD_p from the pilot is:

$$\sqrt{\frac{2(n-1)}{\chi_{0.8,2(n-1)}}}. \quad (5)$$

To consider the impact on power and planned sample size, we need to state reasonable specific alternative hypotheses. In trials, it is uncommon to see large differences between treatments so we considered small to medium standardised effect sizes (differences between the group means) of 0.2, 0.35 and 0.5 [30]. For each true effect size of 0.2, 0.35 or 0.5, we divide by the SD_p estimate for each replicate, and use this value to calculate the required sample size. For each simulated pilot study, we calculate the planned sample size for the RCT assuming either the unadjusted or adjusted SD_p estimated from the pilot. Using this planned sample size (where the SD_p has been estimated) we then calculate the true power of the planned study assuming that we know that the true population SD_p is in fact 1.

Binary outcomes

We consider that the binary outcome will be measured for one homogeneous group only. The following is repeated for each true population success probability. We examined nine true success probabilities from 0.1 to 0.5 in intervals of 0.05. We considered 41 different pilot study sizes ranging from 10 to 200 consisting of multiples of five subjects. The subscripts i and j are used to denote the true proportion and the pilot study size, respectively. For each simulated pilot study of size n_j, the number of successes $(Y_{ij} \sim \mathrm{Bin}(n_j, \theta_i))$ in the simulation n_j are counted. First, the observed proportions, $\hat{\theta}_i$, for each of the nine true success probabilities were calculated by:

$$\hat{\theta}_i = \frac{Y_{ij}}{n_j}. \quad (6)$$

The associated 95% confidence interval was calculated using Wilson's score [21] given by:

$$\frac{\left(\hat{\theta}_i + \frac{z^2_{\alpha/2}}{2n_j} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}_i\left(1-\hat{\theta}_i\right)+\frac{z^2_{\alpha/2}}{4n_j}}{n_j}}\right)}{\left(1+\frac{z^2_{\alpha/2}}{n_j}\right)} \quad (7)$$

Second, this process was repeated for N_s (the number of simulations needed to estimate the true success probability to within 0.1% of its standard error) and the average observed success probability for each of the nine true success probabilities (θ) for a given fixed pilot size were calculated as follows:

$$\bar{\theta}_i = \frac{1}{N_s}\sum_{k=1}^{N_s}\hat{\theta}_{ik}, \quad (8)$$

where $\hat{\theta}_{ik}$ is $\hat{\theta}_i$ for the kth simulated pilot study. Third, due to the relatively small sample size of the pilot trials, we computed the mean width of the 95% confidence interval of the true success probability averaged over N_s simulations using the Wilson's score method [31] for a fixed sample size, which is given by:

$$\frac{1}{N_s}\sum_{k=1}^{N_s}\frac{\left(2z_{a/2}\sqrt{\frac{\hat{\theta}_{ik}\left(1-\hat{\theta}_{ik}\right)+\frac{z^2_{a/2}}{4n_j}}{n_j}}\right)}{\left(1+\frac{z^2_{a/2}}{n_j}\right)}. \quad (9)$$

The relative percentage gain in precision around the true binomial proportion per increase of five study participants is defined as before:

$$\left(\frac{\mathrm{WCI}_{n_j} - \mathrm{WCI}_{n_j+5}}{\mathrm{WCI}_{n_j}}\right) \times 100. \quad (10)$$

As for the continuous outcomes, bias is assessed by subtracting the true population value from each estimate and taking the signed mean of these. We also report the 95% coverage probability [32].

Results and discussion

Normally distributed outcomes

Figure 1 is a multiple box and whisker plot of the resulting distributions of the sample SD_p. Under our simulations the true SD is equal to 1. Figure 1 clearly shows that the spread of the estimates reduces as the pooled sample size increases and the distribution of the estimated SD_p also becomes more symmetric as the pooled sample size increases. So the bias and skew is more marked for smaller sample sizes. The direction of the bias means that the SD tends to be underestimated. Once the total sample size is above 50 the average bias becomes negligible and is less than 0.005 below the true value. However, what is more noticeable is the large variation in the sampling distribution for the smaller sample sizes and considerable sampling variation remains even with a large sample size.

Figure 2 shows the percentage gain in precision (the width of the confidence interval for the SD_p) when adding ten more participants to the sample (five to each

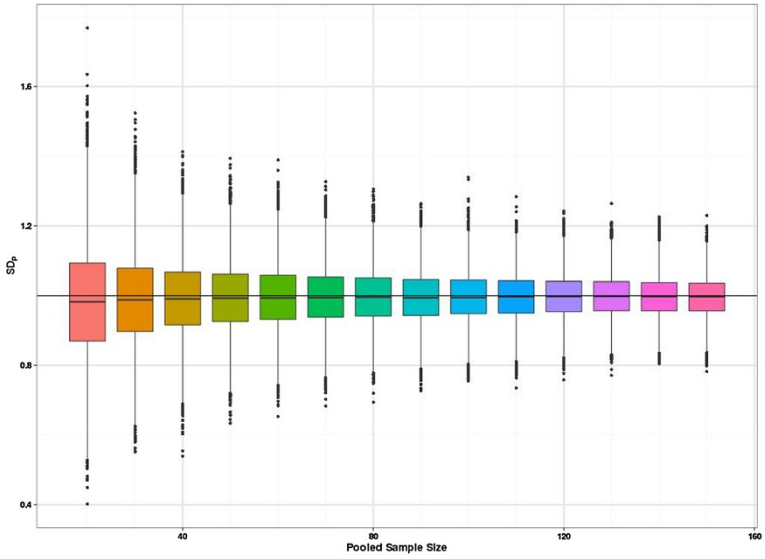


Figure 1 Multiple box and whisker plot of SD_p estimates by pooled sample size of the pilot study. The vertical axis shows the value of the SD_p estimate for 10,000 simulations per pilot study size. The horizontal axis is graduated by the pooled pilot study size.

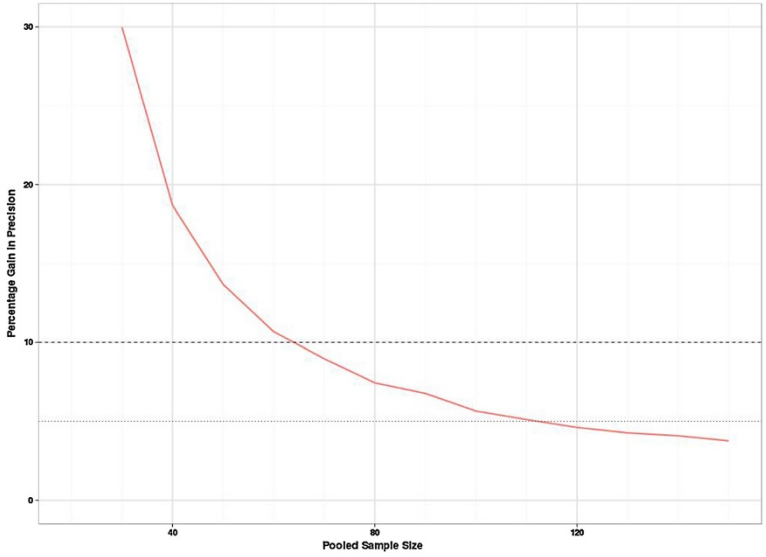


Figure 2 Percentage gain in precision of SD_p on increasing the pooled sample size. This shows the relative reduction in the average width of the confidence interval when an additional five subjects are added to a group.

group). Precision increases with sample size, however, the relative gain in precision (while always positive) decreases with increasing sample size. With a total sample size of 70, there is a less than 10% gain in precision when adding further participants to the study size. So in terms of good precision and minimal bias (for a continuous outcome) a total sample size of 70 seems desirable for a pilot study.

Figure 3 shows the distribution of true power for the planned sample sizes for the specific alternative effect size of 0.2, assuming we require 90% power at the 5% two-sided significance level. The true power distribution for the other effects sizes is very similar (it can be shown that conditional on the estimated SD from the pilot, the distributions should be the same but rounding up to integers causes slight changes at small sample sizes). As anticipated, this figure shows a large variation in power for the smaller sample sizes. However, even with the relatively small pilot sample size of 20, the planned studies do have at least 80% power to detect the target effect

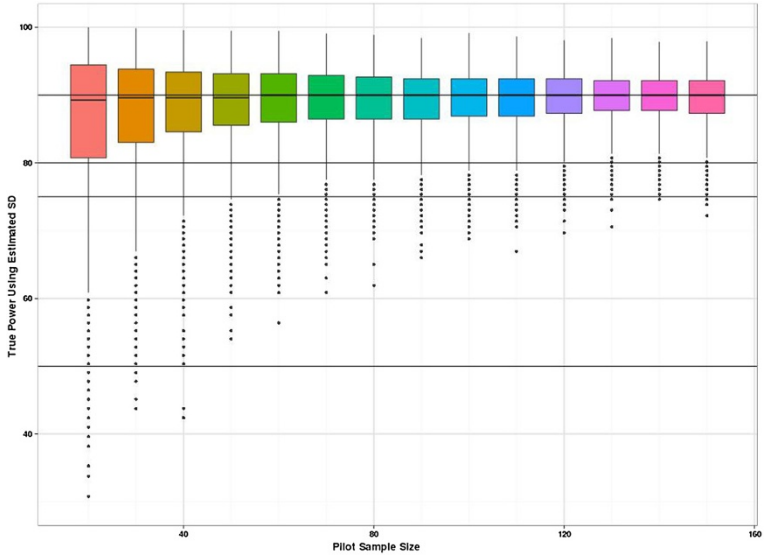


Figure 3 Distribution of planned RCT study power when using the SD_p estimate derived from the pilot study. The planned study size is used to calculate the true power if SD = 1 is assumed. The graph shown is for a true effect size of 0.2. The vertical axis is true power. The *x*-axis shows the size of the two-arm pilot study.

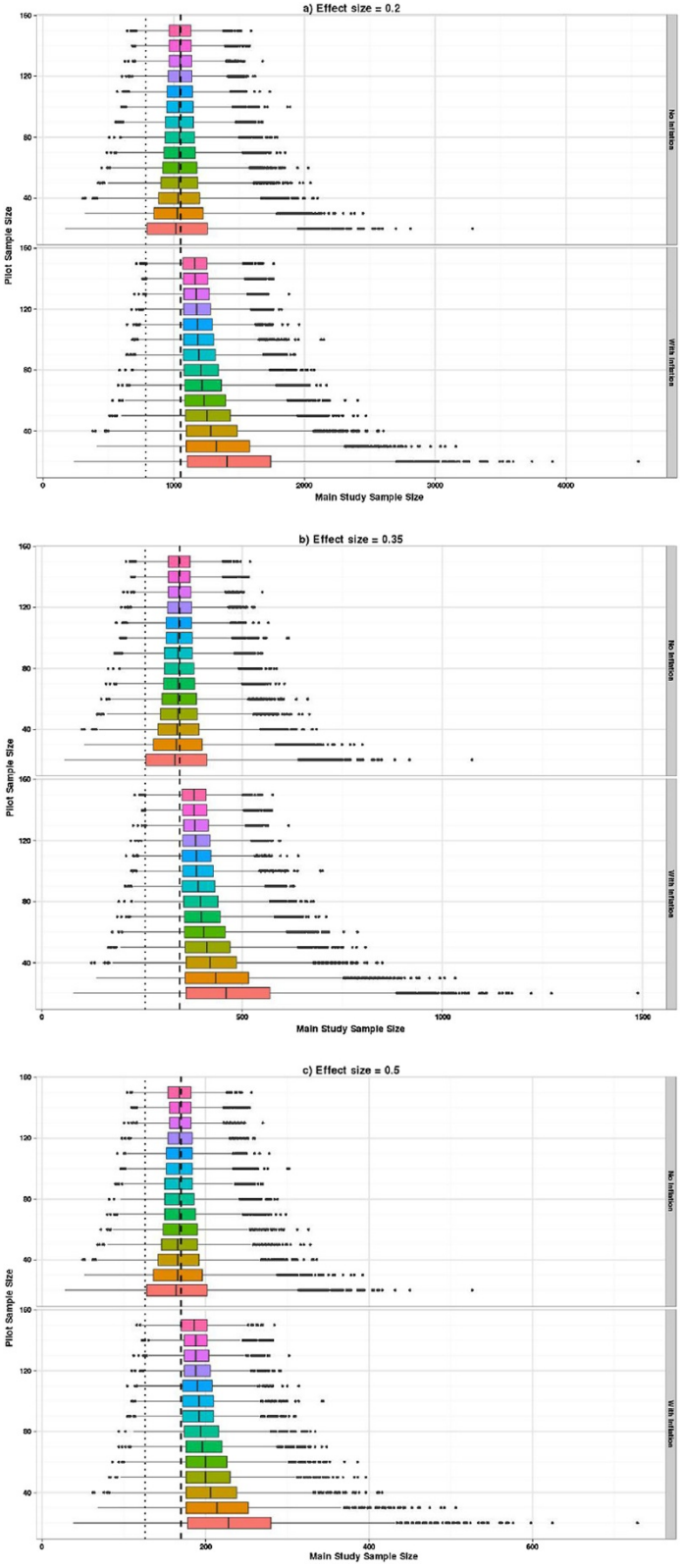


Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Distribution of planned sample sizes using crude SD*p* estimates and adjusting for a specified level of confidence. (a) Effect size = 0.2. **(b)** Effect size = 0.35. **(c)** Effect size = 0.5. The upper part of each graph shows the distribution of planned sample sizes by pilot study size. The lower part shows the same but using the inflation adjustment to guarantee the specified power with 80% confidence. The *x*-axis shows the planned sample size and the vertical axis shows the pilot study size. The dashed vertical line shows the sample size associated with a true power of 90% and the dotted line for 80%.

size (when we have stated we desire 90% power) more than 75% of the time. Figure 3 also shows that the true power frequently exceeds 90% but the cost of this higher power in terms of total participants cannot be quantified from this figure. By contrast Figure 4 is able to show the 'cost' of the higher power translated into the sample size scale.

Figure 4 shows the distribution of the planned sample size when using the estimated SD_p from the pilot (with and without inflation of the SD_p). It can be seen that the overall shape of these plots is similar for all three effects sizes, but the planned sample sizes are proportionately higher as the effect size reduces. Figure 4a shows the sample size (for a true difference between the means of 0.2) using the unadjusted SD_p (upper plot) and the inflated SD_p (lower plot). Using the inflated SD_p means we have specified that we want our planned study to have 90% power with 80% confidence or certainty. By comparing these two plots and superimposing the sample size of 1,052, which is what we would actually need to detect an effect size of 0.2 with 90% power and 5% two-sided significance when the true SD is known to be equal to 1, you can readily see the effect of the inflation factor. Figures 4b,c present the same contrasts as Figure 4a but for a true difference between the means of 0.35 and 0.5, respectively. The main impact of the inflation factor is to guarantee that 80% of the planned studies are in fact *larger* than they need to be, and for the smaller pilots this can be up to 50% larger than necessary. If only the unadjusted crude estimates from the pilot are used to plan the future study, though we aim for at least 50% of studies to be powered at 90%, inspection of the percentiles shows that that the planned sample size delivers at least 80% power with 90% confidence, when a pilot study of at least 70 is used. Researchers need to consider carefully the minimum level of power they are prepared to tolerate for a worst-case scenario when the population variance is overestimated.

Figure 5 adds the size of the pilot study to the planned study size so the distribution of the overall number of subjects required can be seen. The impact of the inflation factor now depends on the true effect size. If we are planning to use the inflation factor then when the effect size is 0.5 a pilot study of around 30 is optimal. However, the same average number of subjects would result using unadjusted estimates from a pilot study of size 70, and this would result in a smaller variation in planned study size. For the effect size of 0.2 then the optimal pilot study size if applying the inflation factor is around 90, but this optimal size still results in larger overall sample sizes than just using unadjusted estimates from pilot studies of size 150.

Binary outcomes

The sampling distribution when estimating a proportion is a function of the true population proportion so it seems unwise to estimate this from a pooled group unless it is a measure independent of treatment group and there is a strong assumption of equality between groups. We have explored the sampling distributions of the proportions in increments of five rather than ten as we allow the possibility that this may be estimated from one arm. As statistical theory predicts the sampling variation is largest when the true proportion is 0.5 and reduces as the true proportion becomes more different from 0.5, we show the results for the two most extreme proportions considered, i.e. 0.1 and 0.5 (Figure 6). When the true proportion is 0.1 the sampling distribution is slightly skewed with a tendency to underestimate the true value even when uneven pilot arm sizes are used. However, when the true proportion is 0.5 there is no systematic bias in under- or overestimating the parameters from the pilot. Most of the fluctuation is due to deriving estimates from a sample size where the true proportion is not a possible outcome (e.g., if the true proportion is 0.5 but the sample size is 25, then the closest you can observe to the true value is 12/25 or 13/25). Once the pilot sample size is 60 or more then these fluctuations settle down. The relative percentage gain in the precision of estimates is formally presented in Figure 7, where the average width of the 95% confidence intervals for the proportion are compared with the average confidence interval width if another five subjects were added to the sample. This relative percentage gain in precision is shown for true proportions 0.1 and 0.5. For the continuous outcomes we suggested a cut-off of 10% as a threshold. For the binary outcomes we use the 5% threshold as we are moving in steps of five rather than ten. The relative percentage gain in the precision graph crosses the 5% threshold when the sample size is 55 to 60 and crosses the 3% threshold when the sample size is 100. Figure 8 shows the coverage probability for five of the true proportions as sample size increases. This shows how frequently the 95% confidence interval contains the true value. This graph shows considerable fluctuations. Once

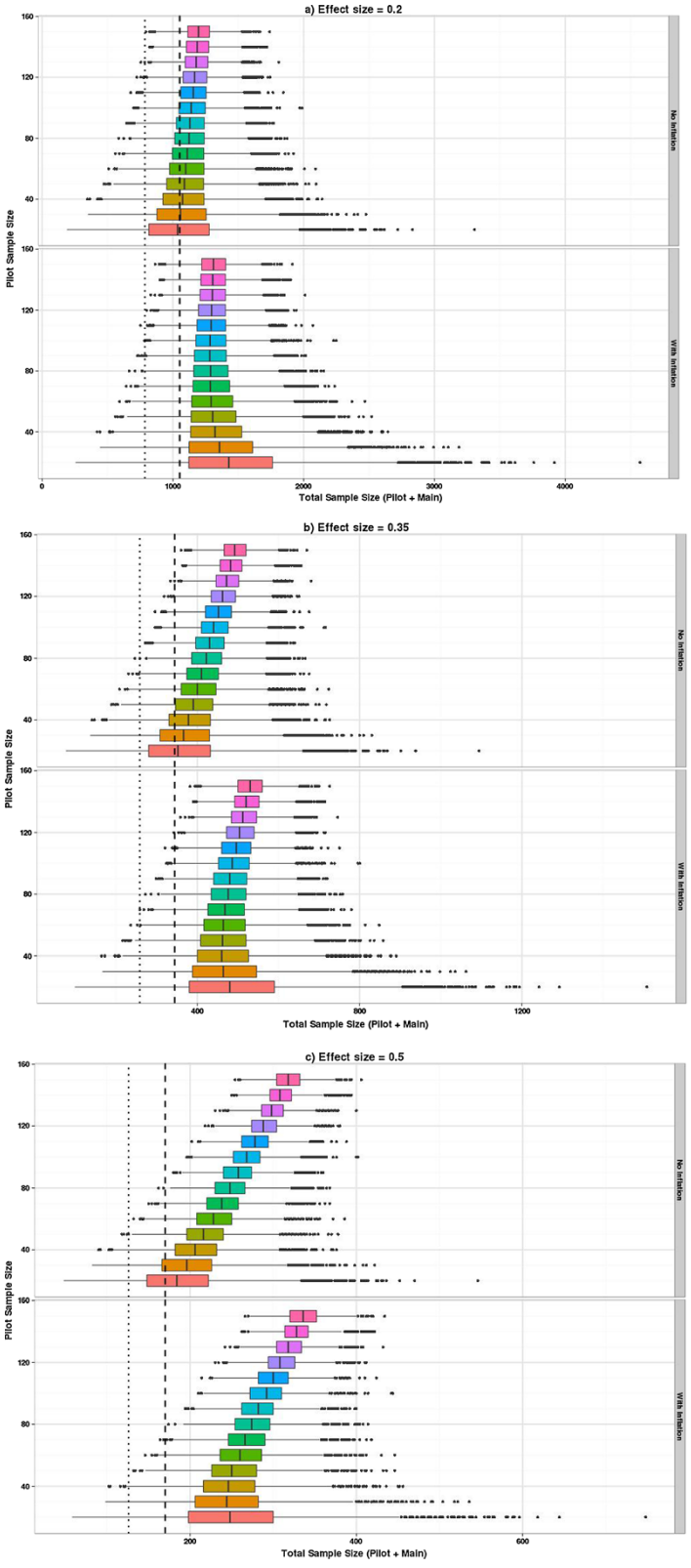


Figure 5 (See legend on next page.)

(See figure on previous page.)
Figure 5 Distribution of total sample size required when using pilot sample derived SD*p* estimated with and without inflation. (a) Effect size = 0.2. **(b)** Effect size = 0.35. **(c)** Effect size = 0.5. This figure is similar to Figure 4; however, now the total sample size includes the pilot study size. The dashed and dotted vertical lines represent the sample size required for 90% and 80% power, respectively, if the true SD were known and the pilot study were not necessary.

the sample size is 100 there is very little perceptible improvement in the coverage probability for the true proportions considered here.

Conclusions

Our simulated data visually demonstrate the large sampling variation that is the main weakness when estimating key parameters from small sample sizes. Small samples sizes do lead to biased estimates, but the bias is negligible compared to the sampling variation. When we examine the relative percentage gain in precision by adding more subjects to the sample, our data suggest that a total of at least 70 may be necessary for estimating the standard deviation of a normally distributed variable with good precision, and 60 to 100 subjects in a single group for estimating an event rate seems reasonable. Treatment-

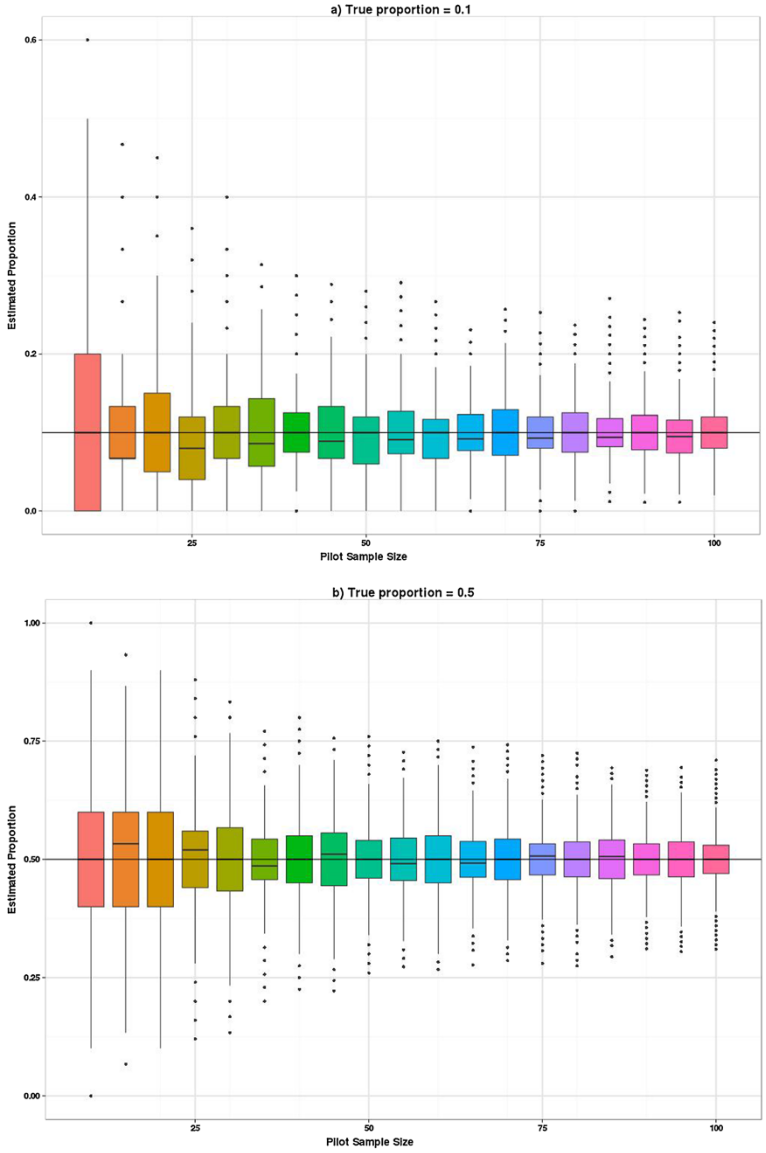


Figure 6 Distribution of estimated event rates on increasing sample size. Distributions for a true event rate of 0.1 **(a)** and a true event rate of 0.5 **(b)**.

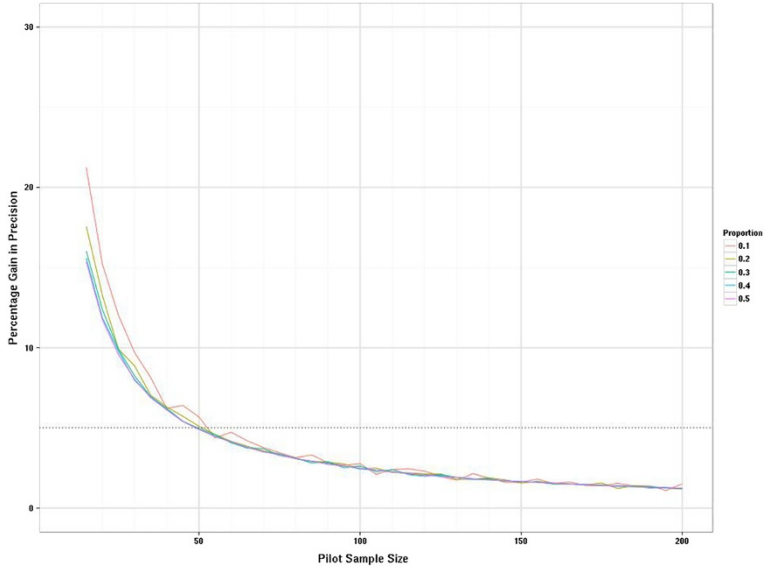


Figure 7 Distribution of relative gain in precision for binary outcomes as pilot study size increases. This graph compares the width of the confidence intervals for $n + 5$ subjects and n subjects. This is scaled by the width of the interval when there are n subjects.

independent parameters may be estimated by pooling the two groups, so in many cases our recommended sample size will be the total sample size. On average when the definitive RCT is planned using an estimate from a pilot study there will be a tendency for the planned study to be underpowered. However, if the definitive RCT is planned for a continuous outcome requiring a power of 90% then the true power will be 80% with at least 76% assurance provided the estimates come from a pilot with at least 20 subjects. We considered three realistic effect sizes of 0.2, 0.35 and 0.5 of a standard deviation to evaluate the impact of adjusting for the anticipated uncertainty in the estimate from the pilot when calculating the sample size for the planned RCT as was recently suggested [18]. For all of the effect sizes considered, it is not efficient to use small pilots and apply the inflation adjustment, as this will result in larger sample sizes (pilot plus main study) in total. Further, we only considered sample sizes planned when requiring 90% power, and examine the conditional power assuming we know the true alternative. On average using imprecise estimates but requiring high power will result in acceptable power with much less 'cost' as measured by total

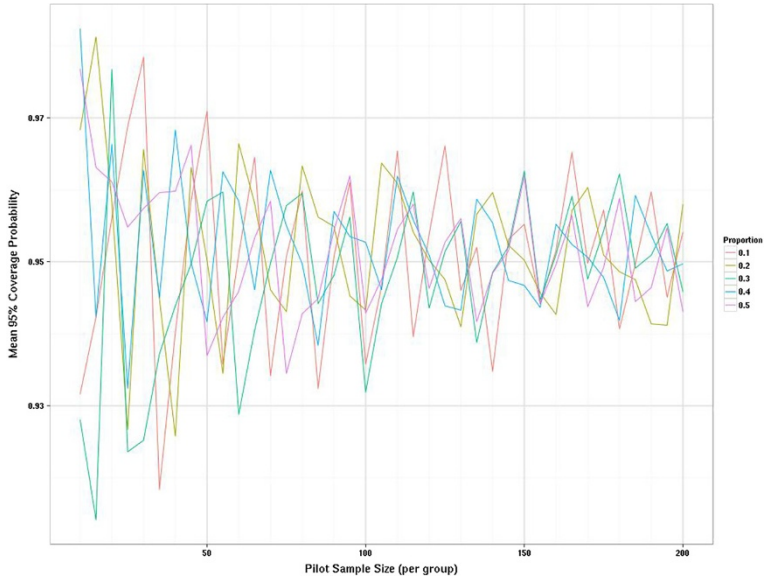


Figure 8 Distribution of mean coverage probability by true proportion and pilot sample size.

sample size. Hence, it is actually more efficient to use a large external pilot study to reduce the variation around the target power for the definitive RCT.

The implication of using estimates of key parameters from small pilot studies is the risk of both over- and underpowered studies. While overpowered studies may not seem such an acute problem, they are potentially a costly mistake and may result in a study being judged as prohibitively large. This would seem to be an argument in favour of utilising internal pilot studies, but an internal pilot requires the key design features of the trial to be fixed, so any change in measurement of the treatment effect following an internal pilot will lead to analysis difficulties.

A major and well-documented problem with published trials is under recruitment, where there is a tendency to recruit fewer subjects than targeted. One reason for under recruitment may well be that event rates such as recruitment and willingness to be randomised cannot be accurately estimated from small pilots, and in fact increasing the pilot size to between 60 and 100 per group may give much more reliable data on the critical recruitment parameters.

In reality, when designing external pilot trials, there is a need to balance two competing issues: maximising the precision (of the critical parameters you wish to estimate) and minimising the size of the external pilot trial, which impacts on resources, time and costs. Thus there is a trade-off between the precision (of the estimates of the critical parameters) and size (number of subjects) of the pilot study. When designing external pilot trials, researchers need to understand that they are trading off the precision of the estimates against the total sample size of the definitive study when they decide to have an external pilot study with a small sample size.

Abbreviations

NICE: National Institute for Health and Care Excellence; NIHR: National Institute for Health Research; RCT: randomised control trial; SD: standard deviation; UK: United Kingdom.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NS, MD and AH contributed to the conceptual design, performed the simulations and summary statistical analysis, and produced the graphical output. MDT contributed to the design of the project, and drafted and revised the manuscript. AW contributed to study design and drafted the literature review. SJW contributed to study design, and the first draft and revisions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

MDT, SJW, AW and NS are funded by the University of Sheffield. MD is fully funded by NIHR as part of a doctoral research fellowship (DRF-2012-05-182). AH was funded by NIHR-Research Design Service and the University of Sheffield. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health or organisations affiliated to or funding them.

The authors thank the three reviewers for their detailed critical comments, which substantially improved the manuscript. We also thank members of the Medical Statistics Group at the School of Health and Related Research, University of Sheffield, for constructive discussions and input to the project. We acknowledge the University of Sheffield for supporting this research.

Received: 13 December 2013 Accepted: 20 June 2014
Published: 3 July 2014

References

1. NIHR Annual Report 2012/2013. [www.nihr.ac.uk/publications]
2. Vickers AJ: **Underpowering in randomized trials reporting a sample size calculation.** *J Clin Epidemiol* 2003, **56**(8):717–720.
3. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P: **Reporting of sample size calculation in randomised controlled trials: review.** *BMJ* 2009, **338:**b1732.
4. Clark T, Berger U, Mansmann U: **Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review.** *BMJ* 2013, **346:**f1135.
5. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, Elbourne DR, Francis D, Garcia J, Roberts I: **What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies.** *Trials* 2006, **7**(1):9.
6. Sully BG, Julious SA, Nicholl J: **A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies.** *Trials* 2013, **14**(1):166.
7. NIHR, Feasibility and pilot studies. [http://www.nets.nihr.ac.uk/glossary]
8. Arnold DM, Burns KEA, Adhikari NKJ, Kho ME, Meade MO, Cook DJ: **The design and interpretation of pilot trials in clinical research in critical care.** *Crit Care Med* 2009, **37**(1):S69–S74.
9. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios L, Robson R, Thabane M, Giangregorio L, Goldsmith C: **A tutorial on pilot studies: the what, why and how.** *BMC Med Res Methodol* 2010, **10**(1):1.
10. Lee EC, Whitehead AL, Jacques RM, Julious SA: **The statistical interpretation of pilot trials: should significance thresholds be reconsidered?** *BMC Med Res Methodol* 2014, **14:**41.
11. Proschan MA: **Two-stage sample size re-estimation based on a nuisance parameter: a review.** *J Biopharm Stat* 2005, **15**(4):559–574.
12. Birkett MA, Day SJ: **Internal pilot studies for estimating sample size.** *Stat Med* 1994, **13**(23–24):2455–2463.
13. Wittes J, Brittain E: **The role of internal pilot-studies in increasing the efficiency of clinical-trials.** *Stat Med* 1990, **9**(1–2):65–72.
14. Friede T, Kieser M: **Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation.** *Pharm Stat* 2013, **12**(3):141–146.
15. Browne RH: **On the use of a pilot sample for sample-size determination.** *Stat Med* 1995, **14**(17):1933–1940.
16. Julious SA: **Sample size of 12 per group rule of thumb for a pilot study.** *Pharm Stat* 2005, **4**(4):287–291.
17. Julious SA: **Designing clinical trials with uncertain estimates of variability.** *Pharm Stat* 2004, **3**(4):261–268.
18. Sim J, Lewis M: **The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency.** *J Clin Epidemiol* 2012, **65**(3):301–308.
19. Kieser M, Wassmer G: **On the use of the upper confidence limit for the variance from a pilot sample for sample size determination.** *Biom J* 1996, **38**(8):941–949.
20. Bland JM: **The tyranny of power: is there a better way to calculate sample size?** *BMJ* 2009, **339:**b3985.
21. Sahu SK, Smith TMF: **A Bayesian method of sample size determination with practical applications.** *J R Stat Soc Ser A – Stat Soc* 2006, **169:**235–253.
22. O'Hagan A, Stevens JW, Campbell MJ: **Assurance in clinical trial design.** *Pharm Stat* 2005, **4**(3):187–201.
23. Brutti P, De Santis F: **Robust Bayesian sample size determination for avoiding the range of equivalence in clinical trials.** *J Stat Plann Inference* 2008, **138**(6):1577–1591.
24. Kirkwood BR, Sterne JAC: *Essential Medical Statistics.* 2nd edition. Oxford: Blackwell Science; 2003.
25. Campbell MJ, Walters SJ, Machin D: *Medical Statistics: A Textbook for the Health Sciences.* 4th edition. Chichester: Wiley; 2007.

26. Satterthwaite FE: **An approximate distribution of estimates of variance components.** *Biometrics Bull* 1946, **2:**110–114.
27. Welch BL: **The generalization of 'Student's' problem when several different population variances are involved.** *Biometrika* 1947, **34:**28–35.
28. StataCorp: *Statistical Software: Release 12.* TX: College Station; 2011.
29. Team RC: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.
30. Cohen J: *Statistical Power Analysis for the Behavioural Sciences.* 2nd edition. Hillsdale, NJ: Lawrence Erlbaum; 1988.
31. Agresti A, Coull BA: **Approximate is better than 'exact' for interval estimation of binomial proportions.** *Am Statistician* 1998, **52**(2):119–126.
32. Burton A, Altman DG, Royston P, Holder RL: **The design of simulation studies in medical statistics.** *Stat Med* 2006, **30**(25):4279–4292.

doi:10.1186/1745-6215-15-264
Cite this article as: Teare *et al.*: Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014 **15**:264.

Lee et al. BMC Medical Research Methodology 2014, 14:41
http://www.biomedcentral.com/1471-2288/14/41

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

The statistical interpretation of pilot trials: should significance thresholds be reconsidered?

Ellen C Lee[†], Amy L Whitehead[†], Richard M Jacques[†] and Steven A Julious[*†]

Abstract

Background: In an evaluation of a new health technology, a pilot trial may be undertaken prior to a trial that makes a definitive assessment of benefit. The objective of pilot studies is to provide sufficient evidence that a larger definitive trial can be undertaken and, at times, to provide a preliminary assessment of benefit.

Methods: We describe significance thresholds, confidence intervals and surrogate markers in the context of pilot studies and how Bayesian methods can be used in pilot trials. We use a worked example to illustrate the issues raised.

Results: We show how significance levels other than the traditional 5% should be considered to provide preliminary evidence for efficacy and how estimation and confidence intervals should be the focus to provide an estimated range of possible treatment effects. We also illustrate how Bayesian methods could also assist in the early assessment of a health technology.

Conclusions: We recommend that in pilot trials the focus should be on descriptive statistics and estimation, using confidence intervals, rather than formal hypothesis testing and that confidence intervals other than 95% confidence intervals, such as 85% or 75%, be used for the estimation. The confidence interval should then be interpreted with regards to the minimum clinically important difference. We also recommend that Bayesian methods be used to assist in the interpretation of pilot trials. Surrogate endpoints can also be used in pilot trials but they must reliably predict the overall effect on the clinical outcome.

Keywords: Pilot trial, Power, Type I error, Confidence interval, Significance, Bayesian methods

Background

In an evaluation of a new health technology, a pilot trial may be undertaken prior to a definitive trial that makes a definitive assessment of benefit. The main objective of a pilot trial is to provide sufficient assurance to enable a larger definitive trial to be undertaken. For example, they may assess aspects such as recruitment rates or whether the technologies can be implemented.

Pilot studies are more about learning than confirming: they are not designed to formally assess evidence of benefit. As such, for clinical endpoints, rather than formal hypothesis testing to prove definitively there is a response, it is usually more informative to provide an estimate of the range of possible responses [1,2]. This estimation may not be around the primary endpoint for the definitive study but could be on a surrogate or an early assessment of an endpoint which may be assessed at a later time point in the definitive study [3].

In this paper we present and discuss approaches towards significance thresholds and confidence interval levels in pilot studies. The methods are divided into three main sections. In the first, we provide alternatives to hypothesis testing using the conventional 5% significance level. We then discuss the use of surrogate outcomes in pilot studies. Finally, a Bayesian approach to significant thresholds is introduced. Throughout the paper we use a worked example to provide illustration to the methods discussed.

Methods and results

Significance and confidence levels

Pilot studies are not formally powered to assess effect. However, it may be of interest to calculate confidence intervals to describe the range of effects, even if this is not a conventional 95% confidence interval. In this

* Correspondence: s.a.julious@sheffield.ac.uk
[†]Equal contributors
Medical Statistics Group, School of Health and Related Research (ScHARR), University of Sheffield, 30 Regent Street, Sheffield S1 4DA, UK

section we give a rational for confidence interval estimation and "hypothesis testing" in pilot studies.

Significance levels and power calculations

Pilot studies are usually underpowered to achieve statistical significance at the commonly used 5% level. Despite recommendations that formal significance levels are not provided for pilot studies, [4,5] many still quote and interpret P-values. In a survey of pilot studies published in 2007–8, Arain et al. [6] found that 81% (21/26) of pilot studies performed hypothesis tests in order to comment on the statistical significance of results. If the primary purpose of a pilot study is to provide preliminary evidence of the efficacy of an intervention, then the significance level can be increased for hypothesis testing [7]. Stallard [8] recommends that the design for a phase II trial is based on a one sided Type I error rate of $\alpha = 0.2$. Whilst Schoenfeld [9] proposed a higher type I error rate for preliminary testing in pilot trials; up to a (one sided) $\alpha = 0.25$. In studies other than drug trials, setting and personnel may not be representative of a future main trial: A pilot trial might see a greater treatment difference due to protocol adherence and enthusiasm in the pilot centre, which might not be replicated in a multicentre trial. Nevertheless, the pilot may still be underpowered for a traditional 5% significance threshold.

It should be noted that in the context of a pilot study a Type I error would have a different impact. For a definitive study, a Type I error would mean therapies or health technologies falsely being concluded as beneficial. As such, in this context they would be referred to as societies risk – such that the wish is to have a Type I error as low as possible. For a pilot study the impact of a Type I error is that a definitive study may falsely be undertaken. Although there is a consequence for patients in the trial – being randomised to therapies when there is equipoise – the impact of this false positive error could be in the main on the sponsor or funder i.e. sponsors spend more money and resources on the 'wrong' study that will not result in a true effect/benefit from the new technology.

The aim of a pilot study, therefore, is to inform both the decision whether to conduct a confirmatory study and the design of the larger confirmatory trial. Any interpreted P-values in a pilot study should be with a disclaimer that the study is not adequately powered [10,11]; and while *post hoc* power calculations are possible [11] they are generally not advisable [12]. Instead, estimation and confidence intervals should be used to infer the size and direction of treatment effect.

Confidence intervals

It is recommended in pilot trials that the focus is on descriptive statistics and estimation rather than formal hypothesis testing [4]. A confidence interval for the treatment effect will inform the decision, amongst other factors, whether or not to perform a confirmatory trial. The confidence interval should be interpreted with regards to the minimum clinically important difference (MCID) [12]; this is the difference between treatment groups that is considered to be clinically meaningful, specified *a priori*. If a confidence interval for the treatment difference crosses zero and the MCID, then the results of the pilot study could be considered to be equivocal. There could be no difference between treatments, or there could be a difference larger than the MCID; the results would not preclude either possibility. This approach is superior to formal hypothesis testing as there is insufficient power to test hypotheses, and its focus on the MCID will help inform the main confirmatory trial. Interpreting confidence intervals this way also helps investigators visualise the evidence of effect from the pilot trial.

It is common to report the 95% confidence interval which corresponds to a 5% significance level. In a pilot study, without adequate power, we can consider investigating confidence intervals of different widths to help inform our decision making, these can then be displayed alongside each other to illustrate the strength of preliminary evidence. We suggest setting minimum prior requirement; that the mean treatment difference is above zero, and that a CI of a certain length includes (or is above) the MCID.

Worked example

The Leg Ulcer Study was a randomised controlled trial designed to investigate the relative cost effectiveness of community leg ulcer clinics that use four layer compression bandaging versus usual care provided by district nurses [13,14]. In the trial 233 patients with venous leg ulcers were allocated at random to the intervention (120) or control (113) group. The SF-36 questionnaire was completed at baseline, three and twelve months post randomisation. For this example we investigate the SF-36 General Health (GH) dimension score. The GH dimension is scored on a 0 (poor) to 100 (good health) scale.

We assume that 3 month data for the first 40 patients is the pilot study data. There were 31 individuals with complete 3 month SF-36 GH dimension data (17 in treatment group and 14 in control group).

Note missing data on 22.5% (9/40) patients is quite high and may be considered unacceptable for a main study. In actuality for this trial there was just 14% (29/230) of missing data for the SF-36 data [15]. For our data we may well have observed a randomly high number. If this was a true pilot study then a missing data rate of 22.5% may need some investigation. There are statistical methods for accounting for missing data [16].

However, the only solution to missing data is not to have any. After a pilot study, measures to ensure complete data would need to be investigated to bring the level of missing data to an acceptable level.

We take the minimum clinically important difference to be a 5 point difference in SF-36 GH dimension scores at 3 months post-randomisation; we assume a standard deviation of 20 points. Without seeing the actual trial results, with 40 individuals, there would be 20% power to detect a 5 point or more difference between the groups if it truly existed which is clearly underpowered by conventional standards. Thus, for such a trial it would be more appropriate to estimate possible effects rather than have formal hypothesis tests.

Table 1 displays the results comparing the mean SF-36 GH dimension scores between the home (control) and clinic (intervention) group. The mean difference was found to be 12.8, which is statistically significant at the 10% but not 5% level; there is some evidence of a difference in SF-36 GH dimension between groups. If the significance level was set to 10%, there would be sufficient preliminary evidence of a treatment difference and this would lead onto a full-scale study.

The leg ulcer randomised controlled trial reported in 1998 obtained appropriate ethics committee approvals [14]. The use of the data from this trial for the work presented in this paper has been approved by School of Health and Related Research (University of Sheffield) ethics as secondary analysis of anonymised data.

Figure 1 shows a range of confidence intervals for the mean difference in SF-36 GH scores between the treatment groups. The 95% CI crosses both 0 and the MCID, this gives inconclusive evidence. The 80% and 90% confidence intervals both exclude 0 and cross the MCID, at these levels there is evidence of a treatment difference which is potentially clinically important. A confidence interval of 75% and smaller would be wholly above or equal to the MCID, suggesting at this level that there is a clinically meaningful difference in SF-36 General Health between the groups.

Outcomes

The NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC) describes a pilot study as a smaller version of the main trial, designed to test whether components of the main study can all work together as well as a preliminary assessment of clinical efficacy. This screening function of pilot studies requires a preliminary evaluation of treatments. Therefore, using the definitive clinical endpoint during a pilot trial may not always be viable. There may be times when measuring the clinical endpoint is not efficient [17]. For example, if the clinical endpoint is the five year survival rate, then an assessment of disease progression or tumour shrinkage may be assessed in the pilot. Such endpoints would be used as surrogates for the definitive endpoint. We will now discuss surrogates in more detail [18].

Table 1 Results from the pilot study comparing 3-month SF-36 GH dimension scores

Mean SF-36 GH dimension score			
Clinic (n = 17)	**Home (n = 14)**	**Difference (95% CI)**	**P-value**
68.0 (sd = 17.6)	55.1 (sd = 19.8)	12.8 (−0.8 to 26.6)	0.065

Surrogate endpoints

In the situations described above an investigator may consider using an endpoint other than the clinical endpoint; a surrogate endpoint. ICH E9 [19] defines a surrogate endpoint as

> *'A variable that provides an indirect measurement of effect in situations where direct measurement of clinical effect is not feasible or practical'.*

Using a surrogate endpoint can reduce the required sample size or the duration of the trial compared to using the clinical endpoint. This leads to cost reductions which may be crucial for trial feasibility [18]. For an endpoint to be considered a surrogate the relationship between it and the clinical outcome must be biologically plausible. In addition, the surrogate must have demonstrable prognostic value for the clinical outcome and there must be evidence from clinical trials that treatment effects on the surrogate outcome correspond to treatments effects on the clinical outcome [19].

The risks involved when using surrogate endpoints

When an aim of a pilot study is to estimate design parameters, using a surrogate endpoint may mean we do not get precise estimates. For example, designing the study based on the surrogate may mean having sub optimal information to estimate the variance of the clinical endpoint or an assessment at an earlier time point. This may mean we do not get an accurate estimate of attrition rates.

A surrogate endpoint must reliably predict the overall effect on the clinical outcome [20]. Otherwise it would be possible to wrongly reject effective treatments or take ineffective treatments through to further testing. If a surrogate does predict clinical benefit it could mean treatment benefits can be brought to patients earlier than if clinical outcomes were used and possibly at a lower cost [21].

Worked example revisited

Using the same data set as in the previous example we now look at the 12 month SF-36 general health (GH)

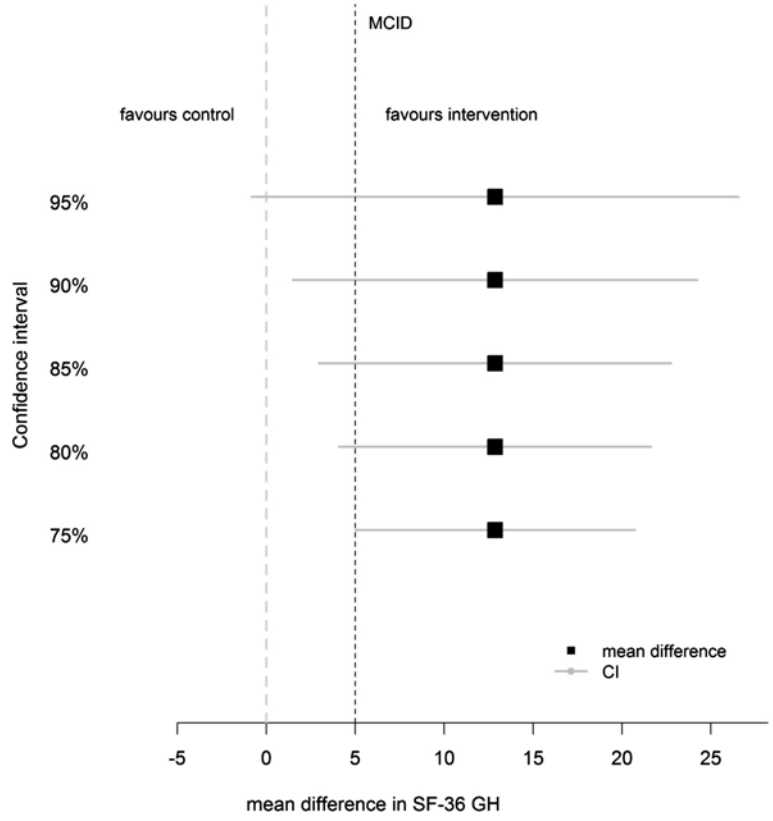


Figure 1 Mean difference in SF-36 GH dimension scores between treatment and control with confidence intervals (based on n = 31 patients).

dimension data for the main trial. There were 233 people in the study in total, 155 with complete SF-36 GH dimension data and 78 observations were recorded as missing. From the 155 observed outcomes 80 were in the clinic group and 75 were in the home or control group – note we had 23% attrition at 3 months compared to 31% at 12 months. Such considerations may be important when trying to design a definitive trial.

Table 2 presents the results from comparing the mean SF-36 GH dimension scores between home and clinic groups. The mean difference was 3.33 which is not significant at the 5% level. The original presentation of these results in 1998 stated that they observed a general deterioration of health status over time, with no difference between the two groups [14].

In the previous worked example we envisaged that the pilot trial had 40 patients and measured the 3-month GH dimension score. Using a significance level of 10% we would have proceeded to the main trial. The 3-month GH dimension score is now considered as a surrogate endpoint to the clinical outcome of 12-month GH dimension score. If we used a significance level of 5% to assess the clinical outcome, the difference between the groups is not statistically significant. Using the 3-month endpoint in the pilot study and a lower significance level would cause us to proceed to the main trial after the pilot study only to observe no significant difference between the two groups in the main study. It could be a Type I error which would lead us to the main study or it could be due to the treatment having no long term efficacy – for example the intervention may have a short term benefit which does not last for 12 months. The 'large' effect of 12.8 points in the first 40 patients at 3 months has not been replicated at 12 months in the full study.

Table 2 Results from main trial comparing 12-month GH dimension scores

Clinic (n = 80)	Home (n = 75)	Difference (95% CI)	P-value
56.0 (sd = 22.8)	52.7 (sd = 23.9)	3.3 (−4.1 to 10.8)	0.377

Mean SF-36 GH dimension score.

Bayesian methods

The Bayesian framework offers an alternative approach to the Frequentist significance levels and confidence intervals discussed in the previous section. It allows prior beliefs about the intervention to be combined with the observed data to form posterior responses about the

outcome of interest. These posterior responses can then be used to inform decisions about whether a larger definitive trial should be undertaken. One approach to making a decision about the intervention is to use a prespecified Go/No-Go criteria.

Go/No-Go criteria

Julious et al. [22] define a Go/No-Go decision as a hurdle in a clinical development path to necessitate further progression or otherwise of a health technology. These hurdles can be set low or high depending on the stage of development of the intervention.

At the planning stage of a pilot study there are a number of decisions that need to be made about how Go/No-Go criteria are defined. The first concerns the metric that is going to measure success or failure. Julious and Swank [23] suggest a method of calculating a probability of success for different development plans based on decision trees and Bayes' Theorem. They take into account the study team's confidence (expressed as a probability) that the intervention will meet the safety and efficacy targets for success, and then calculate the probability that each part of the clinical assessment will correctly indicate that the health technology works or does not work.

Chuang-Stein et al. [24] suggest that a good metric is the probability that there will be a successful confirmatory trial outcome. This is also called assurance by O'Hagan et al. [25] or average power by Chuang-Stein [26] and is used in Bayesian sample size calculations for confirmatory trials. The method that we describe here in detail uses prior beliefs and the data collected from the pilot study to calculate the probability of detecting a clinically meaningful difference. This method has previously been described by Julious et al. [22] for binary and Normal outcomes, and Parmar et al. [27] for survival outcomes.

The second decision concerns the cut-off or level of the criteria. For example, do we want to be 70% or 80% sure that a confirmatory trial will show a minimum clinically meaningful difference? With a pilot study, criteria could be set to minimise the probability of a false positive, (i.e. minimising the probability of progressing an intervention that will fail in a confirmatory trial) but if the goal is set too high then this will increase the probability of a false negative (i.e. stopping an intervention that works from going to a confirmatory trial) [22]. Other factors may also influence the choice of criteria, for example, the sponsor of a drug trial may be more willing to accept an incorrect go decision rather than an incorrect no-go decision if the new treatment is the first in class rather than one of several drugs in class [24].

Prior distributions

As with all Bayesian methods, prior distributions have to be specified for the parameters that we are interested in making inference about and this leads to the question of how these distributions are defined. The simplest approach is to use a non-informative prior. In this case the results will be similar to the Frequentist analysis because all of the information is coming from the observed response. Alternatively, a prior can be elicited based on expert knowledge of the intervention. This may, for example, be based on the synthesis of evidence from previous studies of the same or similar interventions as suggested by Chuang-Stein et al. [24]. Other elicitation techniques including the elicitation from multiple experts are discussed in Spiegelhalter et al. [28].

With a large sample size for the pilot study the posterior distribution will be robust to changes in the prior [29]. However, sample sizes in pilot studies are typically small - in a literature survey by Arain et al. [6] the median number of participants was 76 - and therefore an informative prior distribution may have a large influence on the posterior distribution. We illustrate in our example that caution should be taken when specifying a prior distribution for a pilot study, as different priors may lead to different interpretations of the results.

Probability of detecting a clinically meaningful difference

We now outline one possible method for calculating the probability of detecting a clinically meaningful difference for data that are anticipated to take a Normal form. In the context of a Go/No-Go criteria we need to determine the probability of observing a difference, d_i, or greater given that d_{pilot} has already been observed, i.e. $prob(\theta > d_i \mid d_{pilot})$ where θ is the mean difference.

For Normal data of the form $X_1, X_2, \ldots, X_n \sim N(\theta, \sigma^2)$ we wish to make inference about θ for given σ^2. In this case the Normal family is conjugate and we have the following prior $\theta \sim N(\mu_{prior}, \sigma^2_{prior})$. Note that other distributions may be used for the prior. The Bayesian updating rules can then be defined as follows.

Prior values for the mean difference and population standard deviation are defined as d_{prior} and s_{prior} respectively. The observed mean difference and population standard deviation from the pilot data are defined as d_{pilot} and s_{pilot} respectively. Hence $S_1\sqrt{(r+1)/rn}$ is an estimate of the standard deviation around the mean where r is the allocation ratio between groups and n is the number of individuals per arm.

The posterior distribution is calculated through a weighted sum of the prior and observed responses. The posterior estimate of the mean difference, d_{post}, is defined as

$$d_{post} = s^2_{post}\left(\frac{d_{prior}}{s^2_{prior}} + \frac{d_{pilot}rn}{s^2_{pilot}(r+1)}\right)$$

and the posterior estimate of the variance around the mean, s^2_{post}, is defined as

$$S^2_{post} = \left(\frac{rn}{s^2_{pilot}(r+1)} + \frac{1}{s^2_{prior}} \right)^{-1}.$$

From these posterior values a density distribution for prob($\theta > d_i \mid d_{pilot}$) can be defined so that the probability of observing a difference, d_i, or greater, for a given d_{post} would be

$$prob\left(\theta > d_i | d_{pilot}\right) = \Phi\left(\frac{d_i - d_{post}}{s_{post}}\right).$$

Worked example revisited with bayesian approach

Using the same leg ulcer data as described previously, we demonstrate how to calculate the probability that the mean difference in SF-36 GH dimension scores at 3 months post randomisation is greater than the minimum clinically important difference of five points. This question may also be stated in terms of a 'Go' criteria, for example:

Are we at least 75% sure of having a mean difference in SF-36 GH dimension that is greater than the minimum clinically meaningful difference of five points at 3 months post randomisation.

For the expository purpose of this exercise we will consider the following three Normally distributed priors:

- Non-informative
- Pessimistic prior, with a mean difference of 4 and 90% certainty that the mean difference is within −1 and 9.
- Optimistic prior, with a mean difference of 7 and 90% certainty that the mean difference is within 4 and 10.

Table 3 displays the posterior mean, posterior standard deviation, and the probability that the mean difference in SF-36 GH dimension score is greater than the minimum clinically meaningful difference of 5 points for our examples of a non-informative, pessimistic and optimistic prior distribution. When using both the non-informative and the optimistic prior the probability of achieving a clinically meaningful difference is greater than our pre-set threshold of 75%.

Table 3 Posterior means, standard deviations and the probability of observing a clinically meaningful effect size of greater than 5 for non-informative, pessimistic and optimistic priors

Prior	Posterior mean	Posterior SD	P(>5)
Non-Informative	12.9	6.7	0.88
Pessimistic	5.5	2.8	0.58
Optimistic	7.4	1.8	0.91

Figure 2 shows the prior, observed, and posterior distributions for each of our three examples. The non-informative prior has no influence on the posterior distribution and the 95% credibility interval for the posterior mean difference is

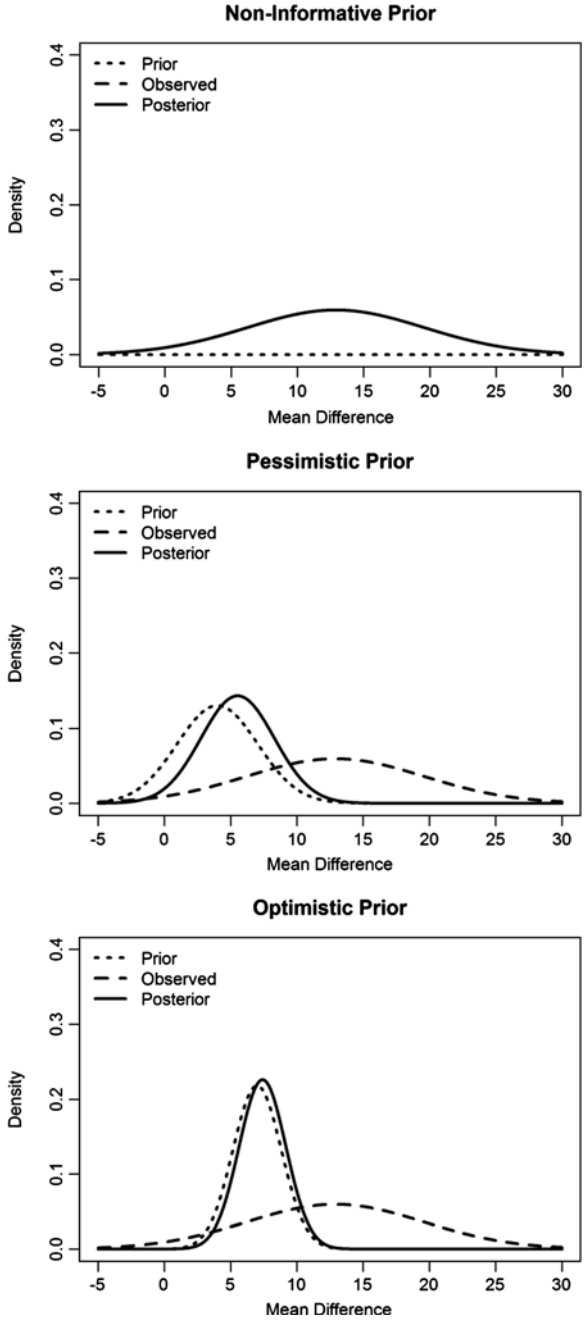


Figure 2 Prior, observed and posterior distributions for non-informative, pessimistic and optimistic priors.

the same as 95% confidence interval found previously (–0.8 to 26.6). In the case of the pessimistic and optimistic priors the posterior distribution is heavily influenced by the choice of prior because the observed data has such a small sample size. This emphasises that caution is required when specifying a prior distribution for pilot studies.

It could be argued that a Bayesian approach is appealing as it formally accounts for any related work (and/or of beliefs held by investigators) by setting priors before the start of a study [22]. Once the trial has been completed, the observed data are combined with the priors to form a posterior distribution for the treatment response. The interpretation is then through a measure that is more easily understood – in our example what is the probability that the response is greater than 5.

Discussion

This paper has demonstrated a variety of approaches towards significance thresholds in pilot studies. When undertaking a pilot investigation, it was shown how significance levels other than the "traditional" 5% should be considered to provide preliminary evidence for efficacy. It was highlighted how estimation and confidence intervals should be focused on in order to provide an estimated range of possible treatment effects.

Interpreting confidence intervals with respect to the minimum clinically important difference should be considered. Investigating several confidence intervals of different widths and displaying them as in Figure 1 can aid decision making and is a helpful way of displaying evidence in pilot studies. Minimum prior requirements can be set and used in addition to the graphical display to help illustrate the strength of preliminary evidence. However, caution must be taken when using a surrogate outcome in pilot studies as it must reliably predict the clinical endpoint.

Bayesian methods could also assist in the early assessment of a health technology. Pilot data can be combined with prior beliefs in order to calculate the probability that there will be a successful confirmatory trial outcome. This can be framed into a Go/No-Go hurdle such as; *are we at least 75% sure of having a mean difference larger than the minimum clinically meaningful difference*. We demonstrated how care must be taken when choosing a prior distribution; the posterior distribution can be heavily influenced by the choice of prior as pilot data usually has a small sample size.

Conclusions

We recommend that in pilot trials the focus should be on descriptive statistics and estimation, using confidence intervals, rather than formal hypothesis testing. We further recommend that confidence intervals in addition to 95% confidence intervals, such as 85% or 75%, be used for the estimation. The confidence interval should then be interpreted with regards to the minimum clinically important difference and we suggest setting minimum prior requirements. Although Bayesian methods could assist in the interpretation of pilot trials, we recommend that they are used with caution due to small sample sizes.

Abbreviations

GH: General Health; MCID: Minimum Clinically Important Difference; NETSCC: National Institute for Health Research Evaluation, Trials and Studies Coordinating Centre.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to the work in this paper. All authors read and approved the final manuscript.

Acknowledgements

We thank Professor Stephen Walters who provided the data used in the worked example. ALW is funded by a School of Health and Related Research (ScHARR) Postgraduate Teaching Assistant Studentship. ECL, RMJ and SAJ did not receive any funding for this work.

Received: 18 October 2013 Accepted: 12 March 2014
Published: 20 March 2014

References

1. Wood J, Lambert M: **Sample size calculations for trials in health services research.** *J Health Serv Res Policy* 1999, **4**(4):226–229.
2. Julious SA, Patterson SD: **Sample sizes for estimation in clinical research.** *Pharm Stat* 2004, **3**(3):213–215.
3. Biomarkers Definitions Working Group: **Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.** *Clin Pharmacol Ther* 2001, **69**(3):89–95.
4. Lancaster GA, Dodd S, Williamson PR: **Design and analysis of pilot studies: recommendations for good practice.** *J Eval Clin Pract* 2004, **10**(2):307–312.
5. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH: **A tutorial on pilot studies: the what, why and how.** *BMC Med Res Methodol* 2010, **10**:1.
6. Arain M, Campbell MJ, Cooper CL, Lancaster GA: **What is a pilot or feasibility study? A review of current practice and editorial policy.** *BMC Med Res Methodol* 2010, **10**:67.
7. Kianifard F, Islam MZ: **A guide to the design and analysis of small clinical studies.** *Pharm Stat* 2011, **10**(4):363–368.
8. Stallard N: **Optimal sample sizes for phase II clinical trials and pilot studies.** *Stat Med* 2012, **31**:1031–1042.
9. Schoenfeld D: **Statistical considerations for pilot-studies.** *Int J Radiat Oncol Biol Phys* 1980, **6**(3):371–374.
10. Papadakis S, Aitken D, Gocan S, Riley D, Laplante MA, Bhatnagar-Bost A, Cousineau D, Simpson D, Edjoc R, Pipe AL, Sharma M, Reid RD: **A randomised controlled pilot study of standardised counselling and cost-free pharmacotherapy for smoking cessation among stroke and TIA patients.** *BMJ Open* 2011, **1**(2):e000366.
11. Legault C, Jennings JM, Katula JA, Dagenbach D, Gaussoin SA, Sink KM, Rapp SR, Rejeski WJ, Shumaker SA, Espeland MA: **Designing clinical trials for assessing the effects of cognitive training and physical activity interventions on cognitive outcomes: the Seniors Health and Activity Research Program Pilot (SHARP-P) study, a randomized controlled trial.** *BMC Geriatr* 2011, **11**:27.
12. Walters SJ: **Consultants' forum: should post hoc sample size calculations be done?** *Pharm Stat* 2009, **8**(2):163–169.
13. Walters SJ, Morrell CJ, Dixon S: **Measuring health-related quality of life in patients with venous leg ulcers.** *Qual Life Res* 1999, **8**(4):327–336.
14. Morrell CJ, Walters SJ, Dixon S, Collins KA, Brereton LML, Peters J, Brooker CGD: **Cost effectiveness of community leg ulcer clinics: randomised controlled trial.** *Br Med J* 1998, **316**(7143):1487–1491.

15. Collins K, Morrell J, Peters J, Walters S, Brooker C, Brereton L: **Problems associated with patient satisfaction surveys.** *Bri J Commun Health Nurs* 2007, **2**(3):156–163.
16. Carpenter JR, Kenward MG: *Multiple Imputation and its Application.* Chichester: Wiley; 2013.
17. De Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL: **Considerations in the evaluation of surrogate endpoints in clinical trials: Summary of a National Institutes of Health Workshop.** *Control Clin Trials* 2001, **22**(5):485–502.
18. Prentice RL: **Surrogate endpoints in clinical-trials - definition and operational criteria.** *Stat Med* 1989, **8**(4):431–440.
19. International Conference on Harmonisation: **ICH E9 statistical principals for clinical trials.** 1998. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.
20. Fleming TR, DeMets DL: **Surrogate end points in clinical trials: are we being misled?** *Ann Intern Med* 1996, **125**(7):605–613.
21. Temple R: **Are surrogate markers adequate to assess cardiovascular disease drugs?** *J Am Med Assoc* 1999, **282**(8):790–795.
22. Julious SA, Machin D, Tan SB: *An Introduction to Statistics in Early Phase Trials.* Oxford: Wiley-Blackwell; 2010.
23. Julious SA, Swank DJ: **Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan.** *Pharm Stat* 2005, **4**(1):37–46.
24. Chuang-Stein C, Kirby S, French J, Kowalski K, Marshall S, Smith MK, Bycott P, Beltangady M: **A quantitative approach for making go/no-go decisions in drug development.** *Drug Inform J* 2011, **45**(2):187–202.
25. O'Hagan A, Stevens JW, Campbell MJ: **Assurance in clinical trial design.** *Pharm Stat* 2005, **4**(3):187–201.
26. Chuang-Stein C: **Sample size and the probability of a successful trial.** *Pharm Stat* 2006, **5**(4):305–309.
27. Parmar MKB, Ungerleider RS, Simon R: **Assessing whether to perform a confirmatory randomized clinical trial.** *J Natl Canc Inst* 1996, **88**(22):1645–1651.
28. Spiegelhalter DJ, Abrams KR, Myles JP: *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Chichester: John Wiley & Sons; 2004.
29. Lee PM: *Bayesian Statistics: An Introduction.* New York: Oxford University Press; Edward Arnold; 1989.

doi:10.1186/1471-2288-14-41
Cite this article as: Lee *et al.*: The statistical interpretation of pilot trials: should significance thresholds be reconsidered?. *BMC Medical Research Methodology* 2014 **14**:41.

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database

Sophie AM Billingham[1], Amy L Whitehead[2] and Steven A Julious[2*]

Abstract

Background: There is little published guidance as to the sample size required for a pilot or feasibility trial despite the fact that a sample size justification is a key element in the design of a trial. A sample size justification should give the minimum number of participants needed in order to meet the objectives of the trial. This paper seeks to describe the target sample sizes set for pilot and feasibility randomised controlled trials, currently running within the United Kingdom.

Methods: Data were gathered from the United Kingdom Clinical Research Network (UKCRN) database using the search terms 'pilot' and 'feasibility'. From this search 513 studies were assessed for eligibility of which 79 met the inclusion criteria. Where the data summary on the UKCRN Database was incomplete, data were also gathered from: the International Standardised Randomised Controlled Trial Number (ISRCTN) register; the clinicaltrials.gov website and the website of the funders. For 62 of the trials, it was necessary to contact members of the research team by email to ensure completeness.

Results: Of the 79 trials analysed, 50 (63.3%) were labelled as pilot trials, 25 (31.6%) feasibility and 14 were described as both pilot and feasibility trials. The majority had two arms (n = 68, 86.1%) and the two most common endpoints were continuous (n = 45, 57.0%) and dichotomous (n = 31, 39.2%). Pilot trials were found to have a smaller sample size per arm (median = 30, range = 8 to 114 participants) than feasibility trials (median = 36, range = 10 to 300 participants). By type of endpoint, across feasibility and pilot trials, the median sample size per arm was 36 (range = 10 to 300 participants) for trials with a dichotomous endpoint and 30 (range = 8 to 114 participants) for trials with a continuous endpoint. Publicly funded pilot trials appear to be larger than industry funded pilot trials: median sample sizes of 33 (range = 15 to 114 participants) and 25 (range = 8 to 100 participants) respectively.

Conclusion: All studies should have a sample size justification. Not all studies however need to have a sample size calculation. For pilot and feasibility trials, while a sample size justification is important, a formal sample size calculation may not be appropriate. The results in this paper describe the observed sample sizes in feasibility and pilot randomised controlled trials on the UKCRN Database.

Keywords: Pilot, Feasibility, Sample size, UK

* Correspondence: s.a.julious@sheffield.ac.uk
[2]Medical Statistics Group, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, Regent Street, Sheffield S1 4DA, UK
Full list of author information is available at the end of the article

Background

The National Institute of Health Research Evaluation, Trials and Studies Coordinating Centre (NETSCC) defines a pilot trial for a randomised controlled trial (RCT) as *'a version of the main study…run in miniature to test whether the components of the study can all work together'* and a feasibility study for an RCT as *'research done before a main study to answer the question "Can this study be done?"'*. [1] However, whilst some authors, including Arain et al. [2] recommend these definitions, in truth there is no consensus. Stallard [3] reports a reason for this as being in part, due to the wide variety of purposes for which pilot trials are undertaken.

Thabane et al. [4] give a number of reasons as to why pilot trials may be conducted. They state that conducting a pilot trial before a main study can increase the likelihood that the main study will be a success, and may potentially help to avoid 'doomed' main trials. They also state that in many cases, pilot trials are performed in order to generate data for sample size calculations in the main study.

Prescott and Soeken [5] meanwhile, suggest five pilot trial aims based on a review of then-current nursing research text books including: a feasibility assessment; adequacy of instrumentation and answering methodological questions.

To address the aims of a pilot trial a sample size justification is required. Hertzog [6] highlights that there is little published guidance on for a pilot trial sample size. However, when applying for funding for a pilot trial, a review panels would expect a justification for the planned sample size. This justification could be based on a number of methods:

- Hertzog [6] recommends the Julious and Patterson [7] method of using confidence intervals for a given precision constructed around the anticipated value to set the sample size;
- Stallard [3] proposes that the sample size should be approximately 0.03 times that the sample size planned to be included in the definitive study;
- Browne [8] gives a general rule is to take a minimum of 30 patients to estimate a parameter;
- Julious [9] recommends a minimum sample size of 12 per group as a rule of thumb and justifies this based on rationale about feasibility and precision about the mean and variance;
- Sim and Lewis [10] suggest a sample size of at least 50 per group.

Setting an appropriate sample size for any study is important. If a study is too large it may be judged to be unethical as participants may be unnecessarily exposed to risks and burdens [11]. There is the additional issue that setting the sample size too high may lead to a preventable failure to reach the recruitment target [12]. While Julious [9] highlights that a sample size that is too small will have an imprecisely estimated variance, which could impact on the design of a future definitive study.

This paper aims to build on the work of Lancaster et al. [12] who reviewed pilot trials published from 2000 to 2001 in seven major journals and Arain et al. [2] who revisited the same seven journals from 2007 to 2008 to see if there had been any change in how pilot trials were reported.

Arain et al. [2] concluded that pilot trials are poorly reported and that the authors are often not explicit as to the purpose of their pilot trial. They also found that sample size calculations were only performed and reported in 35% of the trials and that those identified using the key word 'pilot' were more likely to have a pre-study sample size calculation.

Using data from the United Kingdom Clinical Research Network (UKCRN) Database we extend the work of Lancaster et al. [12] and Arain et al. [2] by investigating the sample size of pilot and feasibility trials for RCTs currently running in the United Kingdom (UK). The aim was to investigate on-going sample sizes for pilot/ feasibility trials in the UK. Although as discussed, there are definitions of pilot and feasibility available, we recognise that in reality the terms are often used interchangeably. However, Arain et al. [2] found that there were some differences between the designs of studies labelled pilot and feasibility. Therefore, in this investigation we will distinguish between pilot and feasibility trials in the analysis. We will further look at whether the sample sizes chosen varies between the two study types (pilot or feasibility), as defined by the principal investigator in their UKCRN Database entry.

The paper will also investigate if the sample size chosen for the trial is influenced by factors such as how the trial is funded or the type of endpoint.

The three research aims of the paper are:

1 To describe the sample sizes set for trials labelled pilot versus feasibility
2 To describe the sample sizes set for trials with a dichotomous compared to a continuous endpoint
3 To describe the sample sizes set in trials funded by industry, public bodies or charities.

Methods

Trial identification

The UKCRN database, [http://public.ukcrn.org.uk/search/ (data last accessed, 20 March 2013)] [13] was used to identify pilot and feasibility trials currently ongoing in the UK. The database comprises of the National Institute for Health Research (NIHR) portfolio in England, and the corresponding portfolios of Northern Ireland, Scotland and

Wales. The studies benefit from the support given by the clinical research network (CRN), however, it is not compulsory for researchers to register with the UKCRN [14]. The database is accessible by anyone online through the URL listed above. The search was conducted on the 17th May 2012 using the key words 'Pilot' or 'Feasibility' in the title or research summary. These were the same key words used by Lancaster et al. [12] and Arain et al. [2] and were used here to maintain consistency with previous research.

The search results were exported to Excel and the studies were sorted first by primary study design in order to separate the interventional trials from the observational studies. They were then sorted by active status: in order to separate the open from the closed trials.

The open interventional trials were then assessed against the eligibility criteria as set out below. After the trials had been assessed against the inclusion criteria the eligible trials were exported into SPSS version 18.0 [15] for analysis.

Trials were eligible for further analysis if:

- They were randomised controlled trials;
- They were currently recruiting participants;
- They were classified as interventional;
- The participants were not healthy volunteers;
- They were not cluster randomised trials.

Trials were only included in the analysis if they were open in order to get the most up to date picture of sample sizes being used for pilot trials in the UK. Trials being conducted on healthy volunteers were not included as these are not usually efficacy studies. Cluster randomised trials were excluded from further analysis as they tend to require much larger target sample sizes (in terms of numbers of patients not clusters) than those trials which randomise patients individually. Cluster randomised trials also have different methodological issues and concerns when undertaking a pilot trial – for example to estimate the intra-class correlation (ICC).

Data extraction

Data on the target sample size and components of the trials that might influence the target sample size such as, type of end point, funder, number of treatment arms and disease area were collected.

The information was extracted from the research summary of the UKCRN database when available. Forty-four of the trials provided an International Standard Randomised Controlled Trial Number [ISRCTN, http://isrctn.org/ (Date last accessed 23rd March 2013)] these were then used to conduct individual searches of the ISRCTN Register, when information was missing.

To complement the search of the UKCRN database, an Internet search was undertaken to find the trial or other websites when information about the trial was missing from the UKCRN. Additional websites used included the US clinicaltrials.gov and the website of the funder of the study.

After conducting all of these searches 62 (75%) of the trials did not have complete information and so, in these cases, the principal investigator or funder(s) were contacted by email for the study protocol in question, in all cases responses were received.

Analysis plan

Medians and ranges were calculated overall for the different types of trial and then broken down by endpoint and whether the trial was public or industry funded.

Results

The search of the UKCRN database yielded 178 studies with the search term 'feasibility' and 335 studies with the search term 'pilot'. After eliminating duplicates, removing any studies not meeting the inclusion criteria and studies where no data were available, 83 trials went on to be analysed. Studies with no data available, means that although the trial was registered, no information regarding the trial was listed or available from other sources. In these cases (n = 5) the trial investigators were contacted however, none of these replied and the trials were assessed as ineligible. Of those eligible, 26 had been labelled as a feasibility by the investigators, 53 had been labelled a pilot trial and 4 had received the label of both a pilot and a feasibility. Figure 1 shows the flow of trials through the review.

Trial characteristics

Table 1 summarises the characteristics of the trials that met the inclusion criteria. The majority of the trials

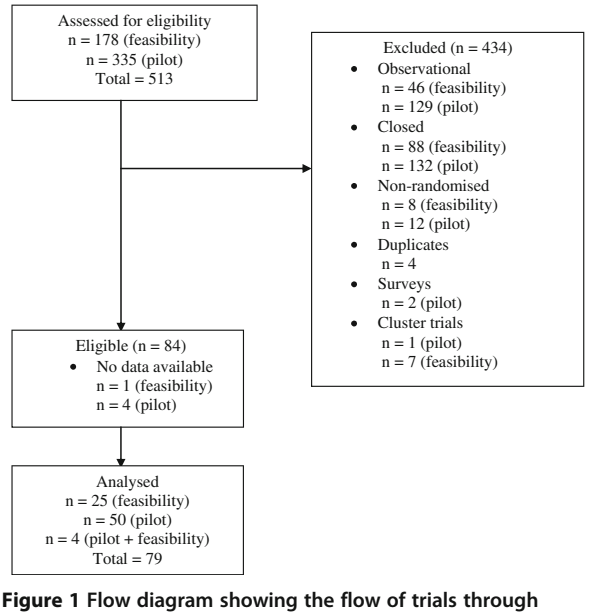


Figure 1 Flow diagram showing the flow of trials through the review.

Table 1 Trial characteristics of the studies included in the final analysis

		Description of preliminary study							
		Pilot		Feasibility		Both		Total	
		n	(%)	n	(%)	n	(%)	n	(%)
Number of arms	Two	39	78.0	25	100.0	4	100.0	68	86.1
	Three	10	20.0	0	0.0	0	0.0	10	12.7
	Four	1	2.0	0	0.0	0	0.0	1	1.3
Type of trial	Health technology	34	68.0	23	92.0	3	75.0	60	75.9
	Drug	16	32.0	2	8.0	1	25.0	19	24.1
Disease area	Stroke	4	8.0	1	4.0	0	0.0	5	6.3
	Mental health	11	22.0	6	24.0	1	25.0	18	22.8
	Oncology	4	8.0	4	16.0	0	0.0	8	10.1
	Respiratory	3	6.0	1	4.0	0	0.0	4	5.1
	Oral & Gastrointestinal	3	6.0	2	8.0	0	0.0	5	6.3
	Dementias	3	6.0	1	4.0	0	0.0	4	5.1
	Cardiovascular	2	4.0	2	8.0	1	25.0	5	6.3
	Primary care	5	10.0	2	8.0	0	0.0	7	8.9
	Musculoskeletal	4	8.0	1	4.0	0	0.0	5	6.3
	Other	11	22.0	5	20.0	2	50.0	18	22.8
Type of end point	Dichotomous	15	30.0	12	48.0	4	100.0	31	39.2
	Continuous	35	70.0	10	40.0	0	0.0	45	57.0
	Time-to-event	0	0.0	1	4.0	0	0.0	1	1.3
	Other	0	0.0	2	8.0	0	0.0	2	2.5
Funder	Industry	11	22.0	1	4.0	1	25.0	13	16.5
	Public	27	54.0	17	68.0	3	75.0	47	59.5
	Charity	12	24.0	7	28.0	0	0.0	19	24.1

(n = 68, 86.1%) consisted of two arms: one experimental treatment and one control treatment, whether that control be active, a placebo or usual care. The majority of the trials had either a continuous endpoint (n = 45, 57.0%) or a dichotomous endpoint (n = 31, 39.2%).

The most common disease areas for the trials were, mental health (n = 18, 22.8%) oncology (n = 8, 10.1%) and primary care (n = 7, 8.9%). Although there was a large variety of clinical areas being investigated as shown in Table 1. Approximately 75% of the trials were health technology trials (n = 60) with drug trials making up the remaining percentage (n = 19).

Most of the trials (n = 47, 59.5%) were publicly funded, with the remaining trials being funded by either a charity (n = 19, 24.1%) or industry (n = 13, 16.5%).

Sample size

The UKCRN database provided a target sample size for each trial in their research summary. However, there were no data available to explain why each target sample size had been chosen.

In approximately 11% of cases (n = 9), the researchers had recruited more patients to date than they initially said would be required. These trials ranged from having a sample size per arm of 15 to 100.

Data were first gathered on the target sample size per arm for pilot and feasibility trials. Those trials labelled pilot were found to have a smaller sample size per arm (median of 30; range 8 to 114 participants) than those labelled feasibility (median of 36; range 10 to 300 participants), these results and the inter-quartile ranges (IQR) are shown in Table 2. Over all, the median sample size per arm was found to be 30 (range 8 to 300).

Data on the median sample size were then analysed according to funder. The results are shown in Table 2. Publicly funded pilot trials have a median sample size of 36 (range 10 to 300 participants) and industry funded pilot trials have a median sample size of 30 (range 8 to 100 participants).

The data were also analysed with regard to type of endpoint used. The results are shown in Table 2. Those studies with a dichotomous endpoint had a median sample size larger than those with a continuous endpoint.

Finally, the data were broken down by both funder and endpoint. The results are shown in Table 3. Public pilot trials with a continuous endpoint were on average

Table 2 Median sample size per arm according to type of study, funder and endpoint

		Sample size per arm		
		n	Median	(IQR) [Range]
Trial description	Pilot	50	30	(20, 45) [8, 114]
	Feasibility	25	36	(25, 50) [10, 300]
	Both	4	49	(36, 61) [23, 72]
Type of endpoint	Dichotomous	31	36	(25, 50) [10, 300]
	Continuous	45	30	(20, 50) [8, 114]
Funder	Industry	13	30	(16, 31) [8, 100]
	Public	47	36	(25, 60) [10, 300]
	Charity	19	30	(20, 45) [15, 52]

larger than industry funded pilot trials with a continuous endpoint (medians of 30 and 23 respectively). The same applies to the public and industry funded pilot trials with a dichotomous endpoint (medians of 36 and 25 respectively). Feasibility trials with a dichotomous endpoint in publicly funded trials are on average larger than the equivalent continuous endpoint trials.

Discussion

Building on the work of Lancaster et al. [12] and Arain et al. [2] the trials analysed in this paper were trials currently running in the United Kingdom on the date the search was conducted, giving us a wide range of information regarding target sample sizes. All the trials that met the inclusion criteria stated a target sample size for their trial within their research summary. Although it is not a requirement in none of the summaries was there a justification given for the target sample size given.

Moore et al. [16] highlighted that it is not unusual for study proposal reviewers to come across a statement such as "No sample size justification is needed because of the pilot nature of the proposed study", but they state that pilot trials are not exempt from needing a clear rationale for the number of patients to be included.

Table 3 Median sample sizes per arm of pilot and feasibility studies by endpoint and funder

			Sample size per arm		
			n	Median	(IQR) [Range]
Pilot	Industry	Dichotomous	5	25	(25, 30) [10, 90]
		Continuous	6	23	(15, 31) [8, 100]
	Public	Dichotomous	6	36	(30, 42) [20, 60]
		Continuous	21	30	(20, 60) [15, 114]
Feasibility	Industry	Dichotomous	0	.	.
		Continuous	1	30	.
	Public	Dichotomous	9	50	(30, 70) [25, 300]
		Continuous	6	43	(15, 60) [10, 60]

However, Arain et al. [2] discovered that only a small proportion of published pilot trials report pre-study sample size calculations as most journal editors state that it is not mandatory criterion for publication.

An investigation of the expected benefits, risks and costs of the study is required to justify a target sample size [16]. However, it is important to remember that a target sample size for a pilot or feasibility study is only a preliminary figure and has a great degree of uncertainty. For example, the researchers may find that more participants drop out than first presumed. We have shown that target sample sizes vary for preliminary trials. Considering the median sample sizes for pilot and feasibility trials our data shows that on average feasibility studies are larger than pilot trials: although there is wide variability in the sample sizes across all types of trial. The median sample size per arm across all the types of study was 30.

With regards to target sample size according to funder, a study of registered drug trials by Bourgeois et al. [17], across a wide variety of types of trial, found that those funded by industry were more likely to have a larger sample size than those funded by government sources. However, our analysis indicated that publicly funded pilot trials were larger than industry funded pilot trials.

Campbell et al. [18] describe sample size calculations for studies that have dichotomous, ordered categorical and continuous endpoints. They state that approximately 30% fewer patients are required for a study with a continuous endpoint – in our research we found that for a dichotomous endpoint compared to a continuous the median sample size was 20% bigger.

Looking at the differences in sample size according to type of primary endpoint and funder we found that there is a larger difference in sample size between trials with a dichotomous endpoint compared to a continuous endpoint for publicly funded trials compared to industry funded trials.

It would be beneficial to follow-up the pilot and feasibility trials discussed in this paper to see how many go on to be published – to see if there is a difference between those published and not published. Another possible extension would be to investigate the different sample sizes of trials dependent on whether the primary endpoint of the trial is based on efficacy or feasibility.

The limitations of this study include the fact that only one trial registry was used to collect the data meaning that it is possible that eligible trials that were not registered with the UKCRN are missing from the analysis. If these trials differ in some way from the trials listed on the UKCRN then this could affect the conclusions made. The database used only trials being carried out in the UK, which could also affect the generalisability of the results. The search was only carried out by one reviewer and was not repeated to check for accuracy. In addition,

only two search terms were used; pilot and feasibility therefore, some trials labelled for example, exploratory or preliminary could have been missed during data extraction. However, these search terms were used to maintain consistency with previous research [2,12].

Conclusion

All trials should have a sample size justification. Not all trials however need to have a sample size calculation. For feasibility and pilot trials, while a sample size justification is important, a formal calculation may not be appropriate. In our study we found that the median pilot study sample sizes for two arm trials were 36 and 30 per arm respectfully for dichotomous and continuous endpoints.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SB identified the trials, extracted the data and performed the analyses. AW created Tables 2, 3 and helped to draft the manuscript. SJ helped to draft the manuscript. All authors read and approved the final manuscript.

Author details

[1]School of Medicine, The University of Sheffield, Beech Hill Road, Sheffield S10 2RX, UK. [2]Medical Statistics Group, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, Regent Street, Sheffield S1 4DA, UK.

Received: 11 November 2012 Accepted: 24 July 2013
Published: 20 August 2013

References

1. *NETSCC definition of pilot and feasibility studies.* [http://www.netscc.ac.uk/glossary/ (date last accessed, 16 August 2013)].
2. Arain M, Campbell MJ, Cooper CL, Lancaster GA: **What is a pilot or feasibility study? a review of current practice and editorial policy.** *BMC Med Res Methodol* 2010, **10:**67. http://www.biomedcentral.com/1471-2288/10/67 (date last accessed 19 August 2013).
3. Stallard N: **Optimal sample sizes for phase II clinical trials and pilot studies.** *Stat Med* 2012, **31:**1031–1042.
4. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, *et al*: **A tutorial on pilot studies: the what, why and how.** *BMC Med Res Methodol* 2010, **10:**1. http://www.biomedcentral.com/1471-2288/10/1 (date last accessed 19 August 2013).
5. Prescott PA, Soeken KL: **The potential uses of pilot work.** *Nurs Res* 1989, **38:**60–62.
6. Hertzog MA: **Considerations in determining sample size for pilot studies.** *Res Nurs Health* 2008, **31:**180–191.
7. Julious SA, Patterson SD: **Sample sizes for estimation in clinical research.** *Pharm Stat* 2004, **3:**213–215.
8. Browne RH: **On the use of a pilot sample for sample size determination.** *Stat Med* 1995, **14:**1933–1940.
9. Julious SA: **Sample size of 12 per group rule of thumb for a pilot study.** *Pharm Stat* 2005, **4:**287–291.
10. Sim J, Lewis M: **The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency.** *J Clin Epidemiol* 2012, **65:**301–308.
11. Altman DG: **Statistics and ethics in medical research III: How large a sample?** *Br Med J* 1980, **281:**1336–1338.
12. Lancaster GA, Dodd S, Williamson PR: **Design and analysis of pilot studies: recommendations for good practice.** *J Eval Clin Pract* 2002, **10**(2)**:**307–312.
13. UKCRN: http://public.ukcrn.org.uk/search/ (date last accessed, 20 March 2013).
14. NIHR: *NIHR clinical research network portfolio.* 2013. [cited 2013 15 March]; Available from: http://www.crncc.nihr.ac.uk/about_us/processes/portfolio/portfolio, [date last accessed 20th March 2013].
15. SPSS Inc: *Released 2009. PASW statistics for windows, version 18.0.* Chicago: SPSS Inc; 2009.
16. Moore CG, Carter RE, Nietert PJ, Stewart PW: **Recommendations for planning pilot studies in clinical and translational research.** *Clin Transl Sci* 2011, **4**(5):332–337.
17. Bourgeois FT, Murthy S, Mandl KD: **Outcome reporting among drug trials registered in ClinicalTrials.gov.** *Ann Intern Med* 2010, **153:**158–166.
18. Campbell MJ, Julious SA, Altman DG: **Sample sizes for dichotomous, ordered categorical and continuous outcomes in two group comparisons.** *Br Med J* 1995, **311:**1145–1148. With Erratrum 1996, **312,** 96.

doi:10.1186/1471-2288-13-104
Cite this article as: Billingham *et al.*: An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Medical Research Methodology* 2013 **13**:104.